Assessment Center Construct-Related Validity:

A Look From Different Angles

Thesis

presented to the Faculty of Arts

of

the University of Zurich

for the degree of Doctor of Philosophy

by

Andreja Wirz-Rodella

of Zurich

Accepted in the spring semester 2012 on the recommendation of

Prof. Dr. Martin Kleinmann and Prof. Dr. Filip Lievens

2012

Contents

Danksagung

Viele Personen haben mich während meiner Arbeit an der Dissertation begleitet und unterstützt. Alle haben auf ihre Weise zum Gelingen dieser Arbeit beigetragen, und dafür ich bin jedem Einzelnen sehr dankbar.

In erster Linie möchte ich mich bei meinem Doktorvater Prof. Dr. Martin Kleinmann und bei Dr. Klaus Melchers bedanken, die sich stets Zeit für meine Anliegen nahmen, mir wegweisende Ratschläge gaben und mich inspirierten. Ihre Unterstützung und ihr engagierter Forschergeist haben diese Arbeit entscheidend geprägt. Besonderer Dank gilt auch Prof. Dr. Filip Lievens für seine äusserst hilfreichen Anregungen, die angenehme Zusammenarbeit sowie für die Übernahme des Zweitgutachtens.

Herzlich danken möchte ich ausserdem Dr. Hubert Annen und Prof. Dr. Wilfried De Corte für ihre Kooperation, die ich sehr schätze. Grosser Dank gilt auch meinen Kolleginnen und Kollegen Irène Calanchina, Maike Debus, Pia Ingold, Dr. Anne Jansen, Prof. Dr. Cornelius König, Natalia Merkulova, Isabelle Odermatt, Dr. Sandra Schumacher, Annika Wilhelmy sowie Dr. Silvan Winkler. Durch ihre Hilfsbereitschaft, fröhliche Art oder inspirierende Gespräche trugen sie alle zu einem sehr angenehmen und produktiven Arbeitsklima bei.

Ferner möchte ich mich bei meinen Studierenden bedanken, deren Einsatz und Unterstützung für mich sehr wertvoll waren: Urs Bettler, Sabrina Engeli, Pascale Gschwend und Stefan Schultheiss.

Zum Schluss möchte ich noch meiner Familie meinen Dank aussprechen. Ihr gebührt ganz besonderer Dank, da sie stets an mich geglaubt und mich in vielerlei Hinsicht unterstützt hat. Insbesondere meinem Mann möchte ich für seine Liebe, seinen Rückhalt und sein Vertrauen in mich, die mir eine wichtige Stütze sind und die diese Arbeit vorantrieben, herzlich danken.

Summary

Assessment Centers (ACs) are a diagnostic tool that serve as a basis for decisions in the context of personnel selection and employee development. In view of the far-reaching consequences that AC ratings can have, it is important that these ratings are accurate. Therefore, we need to understand what AC ratings measure and how the measurement of dimensions, that is, construct-related validity, can be improved.

The aims of this thesis are to contribute to the understanding of the construct-related validity of ACs and to provide practical guidance in this regard. Three studies that offer different perspectives on rating accuracy and AC construct-related validity, respectively, were conducted.

The first study investigated whether increasing assessor team size can compensate for missing assessor expertise (i.e., assessor training and assessor background) and vice versa to improve rating accuracy. On the basis of dimension ratings from a laboratory setting (N = 383), we simulated assessor teams of different sizes. Of the factors considered, assessor training was most effective in improving rating accuracy and it could only partly be compensated for by increasing assessor team size. In contrast, increasing the size of the assessor team could compensate for missing expertise related to assessor background.

In the second study, the effects of exercise similarity on AC construct-related and criterion-related validity were examined simultaneously. Data from a simulated graduate AC (N = 92) revealed that exercise similarity was beneficial for construct-related validity, but that it did not affect criterion-related validity. These results indicate that improvements in one aspect of validity are not always paralleled by improvements in the other aspect of validity.

The third study examined whether relating AC overall dimension ratings to external evaluations of the same dimensions can provide evidence for construct-related validity of ACs. Confirmatory factor analyses of data from three independent samples (Ns = 428, 121, and 92) yielded source factors but no dimension factors in the latent factor structure of AC overall dimension ratings and external dimension ratings. This means that different sources provide different perspectives on candidates' performance, and that AC overall dimension ratings and external dimensions ratings cannot be attributed to the purported dimensions.

Taken as a whole, this thesis looked at AC construct-related validity from different angles. The reported findings contribute to the understanding of rating accuracy and construct-related validity of ACs, respectively. Furthermore, they offer a number of implications for practice and research.

Introduction

Assessment centers (ACs) are a widely used diagnostic tool for personnel selection and employee development. Research has repeatedly demonstrated that ACs are criterion valid, that is, they predict job performance (e.g., Arthur, Day, McNelly, & Edens, 2003; Gaugler, Rosenthal, Thornton, & Bentson, 1987; Hardison & Sackett, 2004; Hermelin, Lievens, & Robertson, 2007). In contrast, findings on internal construct-related validity of ACs indicate that it is unclear whether ACs measure the constructs they are designed to measure (e.g., Melchers, Henggeler, & Kleinmann, 2007; Woehr & Arthur, 2003). This is problematic when considering the far-reaching consequences AC ratings can have (cf. Arthur & Day, 2010).

On the one hand, several explanations for the findings on internal construct-related validity of ACs have been offered (e.g. Lievens, 2009; Sackett & Dreher, 1982). These explanations have guided research on identifying moderators of internal construct-related validity and thus form the basis for attempts to improve internal construct-related validity. On the other hand, some researchers have argued that an external construct-related validation approach might be more appropriate for ACs and, therefore, also more promising for finding evidence for AC construct-related validity (e.g., Neidig & Neidig, 1984; Reilly, Henry, & Smither, 1990; Rupp, Thornton, & Gibbons, 2008).

The general focus of this thesis lies on possible explanations for the findings regarding internal construct-related validity of ACs and attempts to improve internal construct-related validity of ACs. In addition, the external construct-related validation approach as an alternative for determining construct-related validity of ACs is of particular interest. That is, by looking at AC construct-related validity from different angles, this thesis aims to contribute to the understanding of AC construct-related validity and to provide practical guidance in this regard. For this purpose, three studies were conducted.

Thesis Outline

In the *Introduction*, I will briefly describe the AC method before summarizing findings on AC criterion-related validity and construct-related validity. Then, I will offer explanations for the findings on internal construct-related validity and present a brief overview of research attempting to improve this internal construct-related validity. Furthermore, I will address the suggestion to relate AC overall dimension ratings to external evaluations of the same dimensions to examine construct-related validity, that is, the suggestion to use an external construct-related validation approach for ACs instead of an internal one. Finally, to specify the aims of this thesis, I will provide a short outline of the three studies conducted. These studies are then presented in *Chapters 1* to *3*. In the *General Discussion*, I will draw the main conclusions from the studies conducted and deduce implications for practice and directions for future research.

The Assessment Center Method

The AC method is "a procedure to evaluate and develop personnel in terms of attributes or abilities relevant to organizational effectiveness" (Thornton & Rupp, 2006, p. 1). Therefore, ACs are designed to simulate job-related situations that allow observing and evaluating behavior that is critical to job performance. Based on this principle, ACs can be used for different purposes (Thornton & Rupp, 2006): First, ACs can serve as a basis for decisions on selection or promotion of candidates. Second, ACs can provide information about candidates' strengths and weaknesses and thus allow organizations to identify training needs. Third, ACs can be used as a way for candidates to develop new skills and change behavior.

The AC is defined as a standardized procedure to evaluate behavioral dimensions with multiple methods and multiple assessors (Arthur & Day, 2010; International Task Force on

Assessment Center Guidelines, 2009). Thereby, different exercises that are assumed to represent various contextual demands of the target position usually represent the multiple methods (Neidig & Neidig, 1984). Typical exercises are role plays, presentations, in-baskets, and group discussions (Eurich, Krause, Cigularov, & Thornton, 2009; Krause & Thornton, 2009). In the different exercises, several trained assessors observe and evaluate the candidates on predefined dimensions. Examples of frequently used dimensions are communication, problem solving, and organizing and planning (Eurich et al., 2009; Krause & Thornton, 2009). After the completion of all exercises, dimension ratings are pooled in an assessor discussion or they are statistically aggregated (International Task Force on Assessment Center Guidelines, 2009). The final ratings that are derived from assessor discussion or statistical aggregation are then used for personnel decisions and feedback to candidates. It is important that AC ratings are as reliable and accurate as possible so that the AC can appropriately serve the different purposes mentioned above.

AC Criterion-Related Validity

A test has criterion-related validity when the test measures relate significantly to a criterion, for example, to job performance. In the AC domain, various studies have demonstrated evidence for AC criterion-related validity. Meta-analyses found correlations between the overall AC rating and job performance ranging from .26 to .40, indicating that the overall AC rating is predictive of job performance (Arthur et al., 2003; Becker, Höft, Holzenkamp, & Spinath, 2011; Gaugler et al., 1987; Hardison & Sackett, 2004; Hermelin et al., 2007). In addition, meta-analytic findings suggest that ratings of specific dimensions are also criterion valid (Arthur et al., 2003; Meriac, Hoffman, Woehr, & Fleisher, 2008). Moreover, ACs contribute to the prediction of job performance beyond cognitive ability tests

or personality inventories (e.g., Dilchert & Ones, 2009; Krause, Kersting, Heggestad, & Thornton, 2006; Melchers & Annen, 2010; Meriac et al., 2008).

AC Construct-Related Validity

A test is considered construct valid if it measures the constructs it is designed to measure. Concerning AC construct-related validity, many studies cast doubts on whether ACs measure the purported constructs. Typically, correlations between ratings of the same dimension from different exercises are low, which is problematic in terms of convergent validity. In contrast, ratings of different dimension from the same exercise usually correlate substantially, indicating a lack of discriminant validity (cf. Melchers et al., 2007; and Woehr & Arthur, 2003, for meta-analytic results). Furthermore, confirmatory factor analyses usually yield exercise factors in the latent factor structure of AC dimension ratings that account for a substantial proportion of variance in ratings, while dimension factors seem to be a less important source of variance in AC dimension ratings – if they are found at all (e.g., Bowler & Woehr, 2006; Hoffman, Melchers, Blair, Kleinmann, & Ladd, 2011; Lance, Lambert, Gewin, Lievens, & Conway, 2004; Lievens, Dilchert, & Ones, 2009). This pattern of results leads to the conclusion that ACs do not have internal construct-related validity with regard to the targeted dimensions, which is especially problematic when AC ratings are used for identifying candidate's strengths and weaknesses, for example.

Explanations for the Findings on Internal Construct-Related Validity of ACs

Several explanations for the failure to find evidence for internal construct-related validity of ACs have been offered (e.g. Lievens, 2009; Sackett & Dreher, 1982). These explanations form the basis for attempts to improve internal construct-related validity. In the following section, I will describe two explanations that have substantially influenced research

on AC construct-related validity, namely rater bias and situational specificity of candidates' behavior.

Several authors have ascribed the findings with regard to internal construct-related validity of ACs to biases on the side of assessors. In this context, rater bias refers to rating inaccuracies because of difficulties during the AC process (cf. Zedeck, 1986). It has been suggested that assessors provide inaccurate and unreliable ratings such that ACs lack construct-related validity as a consequence. Related to this explanation, the limited cognitive capacity model and the expert model described by Lievens and Klimoski (2001) are of relevance.

According to the limited cognitive capacity model (Lievens & Klimoski, 2001), assessors are not able to meet the high cognitive demands of their task due to limited information processing capacities (e.g., Bycio, Alvares, & Hahn, 1987; Gaugler & Thornton, 1989; Melchers, Kleinmann, & Prinz, 2010; Reilly et al., 1990). Therefore, AC dimension ratings are of impaired reliability and accuracy, which results in poor internal construct-related validity. In line with the limited cognitive capacity model, some studies have provided support for the idea that AC design interventions that are assumed to reduce cognitive demands placed on assessors improve internal construct-related validity. For example, it has been demonstrated that rating accuracy as well as internal construct-related validity are better when assessors have to simultaneously observe a lower compared to a higher number of dimensions during the exercises (Gaugler & Thornton, 1989) or when they observe a lower compared to a higher number of candidates in a group discussion (Melchers et al., 2010). Furthermore, specific tools that should facilitate the rating process, for instance behavioral checklists, also lead to improvements in internal construct-related validity of ACs (Reilly et al., 1990).

The expert model (Lievens & Klimoski, 2001) posits that expert assessors benefit from well-established cognitive structures that facilitate the observation and evaluation of candidates during the AC process and thus enable expert assessors to better cope with the high cognitive demands of the rating task. In contrast, assessors without expertise do not possess such well-established cognitive structures. Therefore, they provide less reliable and less accurate ratings and thus also less construct valid ratings than expert assessors. In line with the expert model, rating accuracy (e.g., Lievens, 2001; Woehr & Huffcutt, 1994) and internal construct-related validity of ACs (e.g., Woehr & Arthur, 2003) has repeatedly been found to improve when assessors gained expertise through assessor training. In addition to assessor training, assessor background also contributes to assessor expertise. For example, meta-analytic findings suggest that psychologists provide more construct valid ratings than managers (Woehr & Arthur, 2003).

Another interpretation of the findings concerning internal construct-related validity of ACs is that exercise variance in AC dimension ratings reflects situational specificity of candidates' behavior (e.g., Lance, Foster, Gentry, & Thoresen, 2004; Lance, Foster, Nemeth, Gentry, & Drollinger, 2007; Lance et al., 2000). As mentioned above, ACs are comprised of different exercises that are designed to represent the variability of contextual demands of the target position (Neidig & Neidig, 1984). Thus, different exercises require different behaviors, and candidates manifest different kinds of behaviors in different exercises (cf. Howard, 2008; Lievens & Conway, 2001; Neidig & Neidig, 1984). That is, candidates' performance is cross-situationally inconsistent, which results in low convergence between ratings of a specific dimension across exercises and in substantial exercise effects. This perspective, that conforms with interactionist theories (e.g., Mischel & Shoda, 1995; Tett & Burnett, 2003; Tett & Guterman, 2000), has repeatedly found support. Specifically, it has been demonstrated that evidence for convergence between ratings (Highhouse & Harris, 1993) and for substantial

dimension variance will more likely be established (Sackett & Harris, 1988; Schneider & Schmitt, 1992) when using exercises that pose similar demands on candidates' behavior.

The rater bias explanation and situational specificity explanation for the findings on internal construct-related validity of ACs are not mutually exclusive, and both have empirical support. Therefore, research based on both explanations is important to gain a better understanding of AC construct-related validity.

An Alternative Approach for Construct-Related Validation of ACs

In the previous paragraph, two explanations for the findings concerning internal construct-related validity of ACs and related research were presented. However, some researchers consider the internal construct-related validation approach to ACs to be inappropriate and, therefore, proposed an external construct-related validation approach (e.g., Neidig & Neidig, 1984; Reilly et al., 1990; Rupp et al., 2008). This alternative approach for construct-related validation of ACs is the subject of the next section.

Referring to the nature of ACs and situational specificity of candidates' behavior, some researchers have argued that different exercises do not capture the same aspects of a dimension (Howard, 2008; Lievens & Conway, 2001; Neidig & Neidig, 1984). Therefore, convergence between dimension ratings from different exercises should not be expected. Furthermore, dimension ratings from single exercises are one-item measures and thus might be unreliable (e.g., Arthur, Day, & Woehr, 2008; Howard, 2008), which might be problematic when trying to find evidence for internal construct-related validity of ACs. Based on these arguments, the use of an external construct-related validation approach for ACs has been proposed (cf. Arthur et al., 2008; Neidig & Neidig, 1984; Reilly et al., 1990; Rupp et al., 2008). Specifically, it has been suggested to integrate dimension ratings from single exercises into overall dimension ratings that are expected to be more reliable than single dimension

ratings. These overall dimension ratings that reflect the overall performance on a dimension should be the focal point for construct-related validation of ACs. To examine whether the overall dimension ratings reflect performance on the purported constructs, it has been proposed to compare overall dimension ratings to evaluations of the same dimensions that stem from sources external to the AC, for example, multisource feedback ratings (Rupp et al., 2008).

On the one hand, a few initial studies that used an external construct-related validation approach for ACs found promising results that strengthen the expectation that AC overall dimension ratings relate to external evaluations of the same dimensions. In particular, these studies found that AC dimension ratings correlate more with conceptually related external measures than with conceptually unrelated external measures (e.g., Shore, Thornton, & Shore, 1990; Thornton, Tziner, Dahan, Clevenger, & Meir, 1997). Such externally assessed measures were cognitive ability measures or personality characteristics, for example. On the other hand, there are also arguments against the expectation that relating AC overall dimension ratings to external evaluations of the same dimensions will provide evidence for AC construct-related validity. For example, in multisource feedback ratings, source variance has been found to dominate over dimension variance, which is problematic in terms of construct-related validity (e.g., Hoffman, Lance, Bynum, & Gentry, 2010). Additionally, because the aforementioned studies that used an external construct-related validity approach have some important limitations, they do not allow conclusions concerning the relation between AC overall dimension ratings and evaluations of the same dimensions obtained external to the AC. Thus, it is unclear whether relating AC overall dimension ratings to external ratings of the same dimensions is successful in providing evidence of AC construct-related validity. More research on this issue is needed.

Aim of the Present Thesis

The aforementioned explanations for the findings on internal construct-related validity of ACs have substantially influenced research in the AC domain. Based on these explanations, moderators of internal construct-related validity were identified and interventions to improve rating accuracy and thus internal construct-related validity were derived from them. However, there are still some important theoretical and practical issues concerning rating accuracy and AC construct-related validity that remain unanswered to date. Therefore, this thesis aimed to clarify some of these issues in order to provide AC users with guidance concerning interventions to improve the accuracy and validity of AC ratings and to contribute to the understanding of AC construct-related validity.

The study presented in *Chapter 1* focused on three AC design factors that affect rating accuracy. Specifically, we examined two AC design factors that are related to assessor expertise and that have been found to influence rating accuracy, namely assessor training and assessor background (e.g., Cardy, Bernardin, Abbott, Senderak, & Taylor, 1987; Lievens, 2001; Woehr & Huffcutt, 1994). The third AC design factor of interest was assessor team size. In line with psychometric theory, ratings that are aggregated across multiple assessors should be more accurate than ratings from single assessors (cf. Cronbach, Gleser, Nanda, & Rajaratnam, 1972). However, assessor expertise and assessor team size are not only related to rating accuracy, but also to AC costs. Therefore, it is important for AC users to know the trade-offs between these moderators in affecting rating accuracy. The aim of this study was to investigate whether increasing assessor team size might compensate for missing assessor expertise with regard to its effectiveness in improving rating accuracy and vice versa. As interventions to improve rating accuracy also lead to improvements in AC construct-related validity (Gaugler & Thornton, 1989; Lievens, 2001; Melchers et al., 2010; Schleicher, Day,

Mayes, & Riggio, 2002), the results from this study are also relevant in terms of AC construct-related validity.

The study presented in *Chapter 2* was based on the unitarian framework of validity that assumes that construct-related and criterion-related validity of a test are closely connected to each other (e.g., Binning & Barrett, 1989; Landy, 1986; Messick, 1995). Of particular interest was whether improvements of one aspect of AC validity are paralleled by improvements in the other aspect of AC validity, which is important from a practical view when implementing interventions to improve one aspect of validity. Therefore, we followed recent calls to investigate AC construct-related and criterion-related validity simultaneously (e.g., Lievens, 2009; Lievens et al., 2009; Melchers & König, 2008; Woehr & Arthur, 2003). Thereby, we focused on a moderator that might have differing effects on AC construct-related and criterion-related validity, namely exercise similarity. Referring to the nature of ACs (cf. Howard, 2008; Neidig & Neidig, 1984) and previous findings (e.g., Schneider & Schmitt, 1992), we argued that exercise similarity is beneficial for AC construct-related validity. In contrast, concerning criterion-related validity, using similar exercises might not be beneficial (cf. Lievens et al., 2009). We aimed to clarify whether exercise similarity indeed has diverging effects on AC construct-related and criterion-related validity.

The main subject in *Chapter 3* was the suggestion to use an external instead of an internal construct-related validation approach to find evidence for AC construct-related validity. Specifically, it was proposed to relate overall dimension ratings from an AC to external evaluations of the same dimensions to determine whether the AC measured the targeted dimensions (cf. Arthur et al., 2008; Neidig & Neidig, 1984; Reilly et al., 1990; Rupp et al., 2008). As mentioned above, there are two competing views regarding whether this approach will yield more promising results with regard to AC construct-related validity than the internal construct-related validation approach. However, previous research does not allow

definite conclusions about which of these two views finds more empirical support. Therefore, we investigated the relation between AC overall dimension ratings and ratings of the same dimensions provided by external sources, thereby taking some important methodological issues into account.

References

Arthur, W., Jr., & Day, E. A. (2010). Assessment centers. In S. Zedeck (Ed.), *APA handbook of industrial and organizational psychology: Selecting and developing members for the organization* (Vol. 2, pp. 205-235). Washington, DC: APA.

Arthur, W., Jr., Day, E. A., McNelly, T. L., & Edens, P. S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology, 56*, 125-154. doi:10.1111/j.1744-6570.2003.tb00146.x

Arthur, W., Jr., Day, E. A., & Woehr, D. J. (2008). Mend it, don't end it: An alternate view of assessment center construct-related validity evidence. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 105-111. doi:10.1111/j.1754-9434.2007.00019.x

Becker, N., Höft, S., Holzenkamp, M., & Spinath, F. M. (2011). The predictive validity of assessment centers in German-speaking regions: A meta-analysis. *Journal of Personnel Psychology, 10*, 61-69. doi:10.1027/1866-5888/a000031

Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology, 74*, 478-494. doi:10.1037/0021-9010.74.3.478

Bowler, M. C., & Woehr, D. J. (2006). A meta-analytic evaluation of the impact of dimension and exercise factors on assessment center ratings. *Journal of Applied Psychology, 91*, 1114-1124. doi:10.1037/0021-9010.91.5.1114

Bycio, P., Alvares, K. M., & Hahn, J. (1987). Situational specificity in assessment center ratings: A confirmatory factor analysis. *Journal of Applied Psychology, 72*, 463-474. doi:10.1037/0021-9010.72.3.463

Cardy, R. L., Bernardin, H. J., Abbott, J. G., Senderak, M. P., & Taylor, K. (1987). The effects of individual performance schemata and dimension familiarization on rating accuracy. *Journal of Occupational and Organizational Psychology, 60*, 197-205. doi:10.1111/j.2044-8325.1987.tb00253.x

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Dilchert, S., & Ones, D. S. (2009). Assessment center dimensions: Individual differences correlates and meta-analytic incremental validity. *International Journal of Selection and Assessment, 17*, 254-270. doi:10.1111/j.1468-2389.2009.00468.x

Eurich, T. L., Krause, D. E., Cigularov, K., & Thornton, G. C., III. (2009). Assessment centers: Current practices in the United States. *Journal of Business and Psychology, 24*, 387-407. doi:10.1007/s10869-009-9123-3

Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., III, & Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology, 72*, 493-511. doi:10.1037/0021-9010.72.3.493

Gaugler, B. B., & Thornton, G. C., III. (1989). Number of assessment center dimensions as a determinant of assessor accuracy. *Journal of Applied Psychology, 74*, 611-618. doi:10.1037/0021-9010.74.4.611

Hardison, C. M., & Sackett, P. R. (2004). *Assessment center criterion related validity: A meta-analytic update*. Unpublished manuscript.

Hermelin, E., Lievens, F., & Robertson, I. T. (2007). The validity of assessment centres for the prediction of supervisory performance ratings: A meta-analysis. *International Journal of Selection and Assessment, 15*, 405-411. doi:10.1111/j.1468-2389.2007.00399.x

Highhouse, S., & Harris, M. M. (1993). The measurement of assessment center situations: Bem's template matching technique for examining exercise similarity. *Journal of Applied Social Psychology, 23*, 140-155. doi:10.1111/j.1559-1816.1993.tb01057.x

Hoffman, B., Lance, C. E., Bynum, B., & Gentry, W. A. (2010). Rater source effects are alive and well after all. *Personnel Psychology, 63*, 119-151. doi:10.1111/j.1744-6570.2009.01164.x

Hoffman, B., Melchers, K. G., Blair, C. A., Kleinmann, M., & Ladd, R. T. (2011). Exercises AND dimensions are the currency of assessment centers. *Personnel Psychology, 64*, 351-395. doi:10.1111/j.1744-6570.2011.01213.x

Howard, A. (2008). Making assessment centers work the way they are supposed to. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 98-104. doi:10.1111/j.1754-9434.2007.00018.x

International Task Force on Assessment Center Guidelines. (2009). Guidelines and ethical considerations for assessment center operations. *International Journal of Selection and Assessment, 17*, 243-253. doi:10.1111/j.1468-2389.2009.00467.x

Krause, D. E., Kersting, M., Heggestad, E. D., & Thornton, G. C., III. (2006). Incremental validity of assessment center ratings over cognitive ability tests: A study at the executive management level. *International Journal of Selection and Assessment, 14*, 360-371. doi:10.1111/j.1468-2389.2006.00357.x

Krause, D. E., & Thornton, G. C., III. (2009). A cross-cultural look at assessment center practices: Survey results from Western Europe and North America. *Applied Psychology: An International Review, 58*, 557-585. doi:10.1111/j.1464-0597.2008.00371.x

Lance, C. E., Foster, M. R., Gentry, W. A., & Thoresen, J. D. (2004). Assessor cognitive processes in an operational assessment center. *Journal of Applied Psychology, 89*, 22-35. doi:10.1037/0021-9010.89.1.22

Lance, C. E., Foster, M. R., Nemeth, Y. M., Gentry, W. A., & Drollinger, S. (2007). Extending the nomological network of assessment center construct validity: Prediction of cross-situationally consistent and specific aspects of assessment center performance. *Human Performance, 20*, 345-362.

Lance, C. E., Lambert, T. A., Gewin, A. G., Lievens, F., & Conway, J. M. (2004). Revised estimates of dimension and exercise variance components in assessment center postexercise dimension ratings. *Journal of Applied Psychology, 89*, 377-385. doi:10.1037/0021-9010.89.2.377

Lance, C. E., Newbolt, W. H., Gatewood, R. D., Foster, M. R., French, N. R., & Smith, D. E. (2000). Assessment center exercise factors represent cross-situational specificity, not method bias. *Human Performance, 13*, 323-353. doi:10.1207/S15327043HUP1304_1

Landy, F. J. (1986). Stamp collecting versus science: Validation as hypothesis testing. *American Psychologist, 41*, 1181-1192. doi:10.1037/0003-066X.41.11.1183

Lievens, F. (2001). Assessor training strategies and their effects on accuracy, interrater reliability, and discriminant validity. *Journal of Applied Psychology, 86*, 255-264. doi:10.1037/0021-9010.86.2.255

Lievens, F. (2009). Assessment centres: A tale about dimensions, exercises, and dancing bears. *European Journal of Work and Organizational Psychology, 18*, 102-121. doi:10.1080/13594320802058997

Lievens, F., & Conway, J. M. (2001). Dimension and exercise variance in assessment center scores: A large-scale evaluation of multitrait-multimethod studies. *Journal of Applied Psychology, 86*, 1202-1222. doi:10.1037/0021-9010.86.6.1202

Lievens, F., Dilchert, S., & Ones, D. S. (2009). The importance of exercise and dimension factors in assessment centers: Simultaneous examinations of construct-related and criterion-related validity. *Human Performance, 22*, 375-390. doi:10.1080/08959280903248310

Lievens, F., & Klimoski, R. J. (2001). Understanding the assessment center process: Where are we now? In C. L. Cooper & I. T. Robertson (Eds.), *International Review of Industrial and Organizational Psychology* (Vol. 16, pp. 245-286). Chichester, UK: Wiley.

Melchers, K. G., & Annen, H. (2010). Officer selection for the Swiss armed forces: An evaluation of validity and fairness issues. *Swiss Journal of Psychology, 69*, 105-115. doi:10.1024/1421-0185/a000012

Melchers, K. G., Henggeler, C., & Kleinmann, M. (2007). Do within-dimension ratings in assessment centers really lead to improved construct validity? A meta-analytic reassessment. *Zeitschrift für Personalpsychologie, 6*, 141-149. doi:10.1026/1617-6391.6.4.141

Melchers, K. G., Kleinmann, M., & Prinz, M. A. (2010). Do assessors have too much on their plates? Rating quality and the number of simultaneously observed candidates in assessment center group discussions. *International Journal of Selection and Assessment, 18*, 329-341. doi:10.1111/j.1468-2389.2010.00516.x

Melchers, K. G., & König, C. J. (2008). It is not yet time to dismiss dimensions in assessment centers. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 125-127. doi:10.1111/j.1754-9434.2007.00023.x

Meriac, J. P., Hoffman, B. J., Woehr, D. J., & Fleisher, M. S. (2008). Further evidence for the validity of assessment center dimensions: A meta-analysis of the incremental criterion-related validity of dimension ratings. *Journal of Applied Psychology, 93*, 1042-1052. doi:10.1037/0021-9010.93.5.1042

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749. doi:10.1037/0003-066X.50.9.741

Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review, 102*, 246-268. doi:10.1037/0033-295X.102.2.246

Neidig, R. D., & Neidig, P. J. (1984). Multiple assessment center exercises and job relatedness. *Journal of Applied Psychology, 69*, 182-186. doi:10.1037/0021-9010.69.1.182

Reilly, R. R., Henry, S., & Smither, J. W. (1990). An examination of the effects of using behavior checklists on the construct validity of assessment center dimensions. *Personnel Psychology, 43*, 71-84. doi:10.1111/j.1744-6570.1990.tb02006.x

Rupp, D. E., Thornton, G. C., III, & Gibbons, A. M. (2008). The construct validity of the assessment center method and usefulness of dimensions as focal constructs. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 116-120. doi:10.1111/j.1754-9434.2007.00021.x

Sackett, P. R., & Dreher, G. F. (1982). Constructs and assessment center dimensions: Some troubling empirical findings. *Journal of Applied Psychology, 67*, 401-410. doi:10.1037/0021-9010.67.4.401

Sackett, P. R., & Harris, M. M. (1988). A further examination of the constructs underlying assessment center ratings. *Journal of Business and Psychology, 3*, 214-229. doi:10.1007/BF01014490

Schleicher, D. J., Day, D. V., Mayes, B. T., & Riggio, R. E. (2002). A new frame for frame-of-reference training: Enhancing the construct validity of assessment centers. *Journal of Applied Psychology, 87*, 735-746. doi:10.1037/0021-9010.87.4.735

Schneider, J. R., & Schmitt, N. (1992). An exercise design approach to understanding assessment center dimension and exercise constructs. *Journal of Applied Psychology, 77*, 32-41. doi:10.1037/0021-9010.77.1.32

Shore, T. H., Thornton, G. C., III, & Shore, L. M. (1990). Construct validity of two categories of assessment center dimension ratings. *Personnel Psychology, 43*, 101-116. doi:10.1111/j.1744-6570.1990.tb02008.x

Tett, R. P., & Burnett, D. D. (2003). A personality trait-based interactionist model of job performance. *Journal of Applied Psychology, 88*, 500-517. doi:10.1037/0021-9010.88.3.500

Tett, R. P., & Guterman, H. A. (2000). Situation trait relevance, trait expression, and cross-situational consistency: Testing a principle of trait activation. *Journal of Research in Personality, 34*, 397-423. doi:10.1006/jrpe.2000.2292

Thornton, G. C., III, & Rupp, D. E. (2006). *Assessment centers in human resource management: Strategies for prediction, diagnosis, and development*. Mahwah, NJ: Lawrence Erlbaum.

Thornton, G. C., III, Tziner, A., Dahan, M., Clevenger, J. P., & Meir, E. (1997). Construct validity of assessment center judgements: Analyses of the behavioral reporting method. *Journal of Social Behavior and Personality, 12*, 109-128.

Woehr, D. J., & Arthur, W., Jr. (2003). The construct-related validity of assessment center ratings: A review and meta-analysis of the role of methodological factors. *Journal of Management, 29*, 231-258. doi:10.1177/014920630302900206

Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology, 67*, 189-205. doi:10.1111/j.2044-8325.1994.tb00562.x

Zedeck, S. (1986). A process analysis of the assessment center method. *Research in Organizational Behavior, 8*, 259-296.

Chapter 1

Trade-Offs Between Assessor Expertise and Assessor Team Size in Affecting Rating Accuracy in Assessment Centers

Andreja Wirz[1], Klaus G. Melchers[1], Filip Lievens[2], Wilfried De Corte[2], and Martin Kleinmann[1]

[1]Universität Zürich, Switzerland; [2]Ghent University, Belgium

Manuscript submitted for publication to *International Journal of Selection and Assessment*

Abstract

We compared the effects of assessor training, assessor background, and assessor team size on rating accuracy in an assessment center exercise. Participants (N = 383) with differing backgrounds were randomly assigned to one of three training conditions and then rated candidates in a sales presentation. With the ratings obtained, we simulated assessor teams of different sizes. Of the three factors, assessor training had the strongest effect on rating accuracy. Furthermore, in most conditions, using larger assessor teams also led to more accurate ratings. However, for untrained assessors, using larger assessor teams could only compensate for missing assessor training when assessors had a psychological background, but not if they were managers. Practical implications and directions for future research are discussed.

Assessment centers (ACs) enjoy popularity in both the private and public sectors, where they play an important role in both personnel selection and employee development. ACs are criterion valid (Arthur, Day, McNelly, & Edens, 2003; Gaugler, Rosenthal, Thornton, & Bentson, 1987; Hardison & Sackett, 2004; Hermelin, Lievens, & Robertson, 2007) and they explain incremental variance in job or training performance over and above other procedures that are easier and cheaper to administer such as cognitive ability tests or personality inventories (e.g., Dilchert & Ones, 2009; Krause, Kersting, Heggestad, & Thornton, 2006; Melchers & Annen, 2010; Meriac, Hoffman, Woehr, & Fleisher, 2008). However, ACs are a relatively expensive selection and assessment technique so an important issue for companies is how to reduce costs for ACs while still ensuring the accuracy of the performance evaluations obtained.

Recent surveys revealed that there is considerable variability in the design and implementation of ACs (cf. Eurich, Krause, Cigularov, & Thornton, 2009; Krause & Thornton, 2009). However, currently only limited empirical evidence is available concerning the potential trade-offs between different design factors that are related to both the costs of ACs and to the accuracy of the performance evaluations from these ACs.

Therefore, in the present research we considered three factors that are of importance in this regard: Assessor training, assessor background, and assessor team size. Assessor training and assessor background are related to the expertise of the assessors, and increasing the size of the assessor team might serve as a potential means to compensate for lack of expertise. However, until now, it remains unknown whether increasing the size of the assessor team is indeed a viable way to improve the accuracy of the evaluations from ACs in comparison to factors related to assessor expertise. Therefore, we evaluated whether increasing the size of the assessor team can compensate for missing expertise, so that AC users can be provided with guidance concerning these issues. Specifically, we aimed to compare the effects of

assessor training, assessor background, and assessor team size on the accuracy of ratings in an AC exercise.

Assessor Expertise

It has been shown meta-analytically that assessor characteristics such as expertise moderate AC validity so that ratings provided by assessors with more expertise have better criterion-related and construct-related validity (e.g., Gaugler et al., 1987; Woehr & Arthur, 2003). According to the expert model (Lievens & Klimoski, 2001), expert assessors benefit from well-established cognitive structures when observing and evaluating candidates, whereas assessors without expertise do not. These well-established structures guide the attention, categorization, integration, and recall of observed behavior and enable expert assessors to better cope with the high cognitive demands of the rating task. Consequently, expert assessors are able to provide more reliable and more accurate ratings than assessors with lower expertise, which results in higher AC validity for the former group. Two important factors that contribute to expertise include assessor training and assessor background. In the following paragraphs, we review research related to these two factors.

Assessor training. Several training approaches for improving rating accuracy have been suggested (cf. Bernardin, Buckley, Tyler, & Wiese, 2000; Woehr & Huffcutt, 1994). For example, behavior observation training (BOT), which is based on the assumption that inaccurate ratings stem from a lack of behavioral information, focuses on the improvement of the observation process (i.e., detection, perception, and recall of relevant behavior). In BOT, assessors are instructed to distinguish between observation and evaluation. Furthermore, BOT stresses the importance of being a good observer, of focusing on actual behavior, and of taking notes on behaviors that are observed. Conversely, the major purpose of frame-of-reference (FOR) training consists of imposing a common performance theory on raters,

thereby establishing a common evaluation standard among assessors. In FOR training, assessors learn to identify relevant behavioral aspects related to the dimensions of interest and to assign observed behavior to the appropriate performance level. Hence, FOR training particularly should foster the correct utilization and evaluation of behavioral cues when providing dimension ratings.

Meta-analytic research has confirmed that rater training in general has a positive effect on rating accuracy (Woehr & Huffcutt, 1994) and on AC validity (Gaugler et al., 1987; Woehr & Arthur, 2003). After BOT or FOR training, assessors provide more accurate ratings than after control training. However, FOR training is more beneficial than BOT because rating accuracy is higher after FOR training than after BOT (Lievens, 2001a; Woehr & Huffcutt, 1994). Compared to untrained assessors, FOR trained assessors not only provide more accurate ratings, but also ratings with better discriminant validity and better criterion-related validity (Schleicher, Day, Mayes, & Riggio, 2002). Research by Schleicher and Day (1998) showed that the improved rating accuracy of FOR trained assessors is particularly due to reduced idiosyncratic representations of candidates' performance. Similarly, Gorman and Rentsch (2009) found that rating accuracy after FOR training was higher, the more closely the assessors' performance theory corresponded to the performance theory taught in the training.

Assessor background. In operational ACs, line managers, HR specialists, and psychologists (Eurich et al., 2009; Krause & Gebert, 2003; Krause & Thornton, 2009) typically serve as assessors. It can be assumed that assessors with different backgrounds have different work experience and, therefore, also have different experience with the performance domain. Zedeck (1986) argued that experienced managers have established schemata of managerial performance that facilitate the evaluation of AC candidates. In line with this, previous performance appraisal research (e.g., Cardy, Bernardin, Abbott, Senderak, & Taylor, 1987; Kozlowski, Kirsch, & Chao, 1986) confirmed that raters with experience in the

performance domain (who thus hold appropriate performance schemata) provide more accurate ratings. For example, when rating managers' performance in appraisal interviews, personnel administrators provided more accurate ratings than MBA students, who, in turn, were more accurate than undergraduates (Cardy et al., 1987). Similarly, in the AC domain, Lievens (2001a) found that managers provided more accurate ratings than psychology students for candidates in an AC exercise (even though the former distinguished less between the dimensions than the latter).

Assessor Team Size

Besides assessor training and assessor background, assessor team size is also expected to be related to rating accuracy in ACs. Specifically, when multiple assessors rate a candidate in an exercise and when ratings from these assessors are aggregated, this should lead to more accurate ratings compared to ratings from single assessors, as the aggregation over multiple measurements is a procedure designed to improve behavioral prediction. That is, aggregation over judges "reduces error of measurement associated with the idiosyncrasies of different judges" (Epstein, 1983, p. 368). More precisely, for aggregated ratings psychometric theory states that error components are divided by the number of assessors, which results in a larger proportion of true variance in comparison to ratings from a single assessor (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). Consequently, the accuracy of aggregated ratings should be higher than the accuracy of non-aggregated ratings from single assessors. Thus, enlarging assessor teams and aggregating their ratings is a potential means to improve rating accuracy in ACs.

Trade-Offs Between Assessor Expertise versus Assessor Team Size

Taken together, it can be assumed that assessor training, assessor background, and assessor team size impact rating accuracy. This means that in the composition of assessor

teams, AC users have to carefully decide (a) whether assessor training should be provided, (b) what background assessors should have, and (c) how many assessors shall constitute an assessor team. Each of these decisions has consequences not only for rating accuracy, but also for AC administration and implementation costs. With regard to assessor training, costs arise across different stages of the training process and for different requirements, such as equipment, facilities, personnel, and material (Noe, 2002). Regarding assessor background, managerial assessors are relatively expensive as compared to (internal) psychologists or HR professionals. This is because the assessors' task is not necessarily part of a manager's job. Hence, in contrast to (internal) psychologists or HR professionals, managers who participate in assessor training or in an AC might invoke indirect extra costs. Hence, assessors' background might influence costs in an AC. Finally, concerning the number of assessors, it is obvious that multiple assessors are more expensive than single assessors. Accordingly, the larger the assessor team, the higher the costs.

Depending on assessor expertise, different numbers of assessors might be needed to reach a particular level of rating accuracy. For example, a larger number of untrained assessors might be able to reach similar rating accuracy as a smaller team of trained assessors. That is, increasing the number of assessors in an assessor team might serve to compensate for missing expertise. Conversely, expertise developed through appropriate assessor training or a specific assessor background might reduce the need for a larger assessor team to ensure rating accuracy. Thus, for AC users, a relevant question is how to weigh assessor expertise against the size of the assessor team so that rating accuracy can be ensured while preventing unnecessary increases in AC costs.

Method

We simulated assessor teams of different sizes and with different expertise on the basis of actual ratings obtained in a study by Lievens (2001a). In his study, Lievens explored the effects of two factors that contribute to assessor expertise (assessor training and assessor background). Therefore, we used Lievens' data and extended the 3 (Assessor training with three levels: BOT, FOR training, and control training) × 2 (Assessor background with two levels: managers and I/O psychology students) design with a third factor, namely assessor team size. More precisely, we determined rating accuracy for single assessors and for teams of two to ten assessors with different expertise and thus assessor team size had ten levels (i.e., 1 to 10 assessors). This led to a 3 × 2 × 10 design.

Sample and Procedure

Data from 390 participants were available. Seven participants were excluded from our analyses because of missing data. Thus, the final sample consisted of 225 advanced Master's level I/O psychology students (130 women and 95 men) and 158 managers (35 women and 123 men). More detailed information on the sample can be found in Lievens (2001a).

Participants were told to assume the role of assessors for the selection of a district sales manager. Then, participants received general information about ACs and a description of the job of the district sales manager and the organization. Afterwards, participants were randomly assigned to one of three training conditions: BOT, FOR training, or control training.

In the BOT condition, participants were taught to distinguish between observation and evaluation, and to improve the processes of observing and recording behavior. At the beginning, participants were instructed to make behavioral instead of nonbehavioral descriptions of candidates' behavior. Then, participants learned to classify behavior into dimensions on the basis of the dimension definitions. Next, the trainer instructed participants

to provide dimension ratings according to the amount of behavioral observations made. Participants practiced recording, classifying, and rating with a videotaped candidate. Afterwards, the behaviors that were used to provide dimension ratings were discussed and discrepancies among ratings were clarified. Finally, participants received feedback pertaining to their ratings.

In FOR training, the aim was to establish a common frame-of-reference for the evaluation of AC candidates. To this end, the trainer presented the definitions of the dimensions and gave examples of normative behaviors for different levels of performance. Afterwards, participants completed a written exercise in which they had to assign behavioral incidents to one of the three dimensions and to one of three performance levels. Then, the answers were discussed and feedback was provided to participants. Next, participants practiced the rating task with a videotaped candidate. The participants' dimension ratings were discussed and discrepancies among ratings were clarified. Finally, the trainer provided participants with feedback regarding their ratings.

Participants in the control condition were told that they were expected to watch videotaped AC candidates, to take notes if necessary, and to evaluate the candidates. Then, participants observed and evaluated a videotaped candidate. However, their ratings were not discussed and no feedback was provided. Hence, participants in the control training did not get a specific preparation for rating AC candidates and thus were untrained.

After BOT, FOR training, or control training, participants observed four videotaped candidates who had to deliver a sales presentation (also see Lievens, 1999, for additional information concerning the development of the videotapes). All participants were unfamiliar with the specific presentation exercise. In this sales presentation, candidates had to present an analysis of the buyer's needs and to argue which of three software systems was most appropriate. The presentation was given to a panel of decision makers who asked questions to

challenge the candidate. The candidates were semiprofessional actors who performed according to pre-specified scripts. The scripts were written on the basis of predefined performances on three dimensions. The predefined performances were later used as true scores to determine rating accuracy as described in the following paragraph. The three dimensions were problem analysis and solving, interpersonal sensitivity, and planning and organization. After every videotaped presentation, each participant had to rate the candidate on the three dimensions using a five-point scale (ranging from 1 = *poor* to 5 = *excellent*). For more details of the procedure we refer to Lievens (2001a).

Rating Accuracy

Using the ratings provided by the participants, we sampled a total of 1000 assessor teams (with replacement after each team was sampled) for all cells of the study design with a size between two to ten assessors. For example, we randomly drew 1000 teams of ten assessors in such a way that each assessor could be sampled in multiple teams. Afterwards, for each of the 1000 teams that consisted of two to ten assessors, we calculated average ratings for the three AC dimensions.

Then, we determined rating accuracy for single assessors and for teams between two to ten assessors. Rating accuracy refers to deviations between the assessor's ratings and comparison scores (cf. Sulsky & Balzer, 1988). Therefore, we compared the dimension ratings obtained with the predefined performances that the scripts for the candidates were based on to determine Borman's differential accuracy (BDA, Borman, 1977). BDA reflects the correlation between ratings and true scores (i.e., the predefined performances that the scripts for the candidates were based on) across candidates, averaged across dimensions. Hence, BDA is a correlational measure of rating accuracy that represents an index of rater validity (Sulsky & Balzer, 1988) and can be expressed by the following equation:

$$BDA = 1/d \sum_{j=1}^{d} (T_{rt})_j \qquad \text{(Equation 1)}$$

where d refers to the number of dimensions and T_{rt} refers to the correlation between ratings r and true scores t for a particular dimension j (Sulsky & Balzer, 1988). Before computing BDA, all correlations T_{rt} are transformed to Z scores.

Results

Effects of Assessor Training, Assessor Background, and Assessor Team Size

The mean rating accuracy for each cell of the study design is presented in Figure 1. To determine the effects of assessor training, assessor background, and assessor team size on rating accuracy, we conducted a three-way analysis of variance (ANOVA) with BDA as the dependent variable. In order to keep the cell sizes balanced, we only used cells with a sample size of 1000 for the analyses, that is, only cells with two to ten assessors per assessor team. This resulted in a 3 × 2 × 9 design for the ANOVA.

In line with psychometric theory, assessor team size had a significant main effect on BDA, $F(8, 53946) = 906.76$, $p < .01$, $\eta^2 = .07$. According to conventional standards (cf. Cohen, 1988), this reflects a moderate effect size. The significant effect of assessor team size indicates that rating accuracy usually increased with an increasing number of assessors in the assessor team (also see Figure 1).

Furthermore, the three-way ANOVA yielded a main effect for assessor training on BDA, $F(2, 53946) = 6824.31$, $p < .01$, $\eta^2 = .14$, with a large effect size (cf. Cohen, 1988). As shown in Figure 1, assessors who had taken part in FOR training provided the most accurate ratings and untrained assessors provided the least accurate ratings.

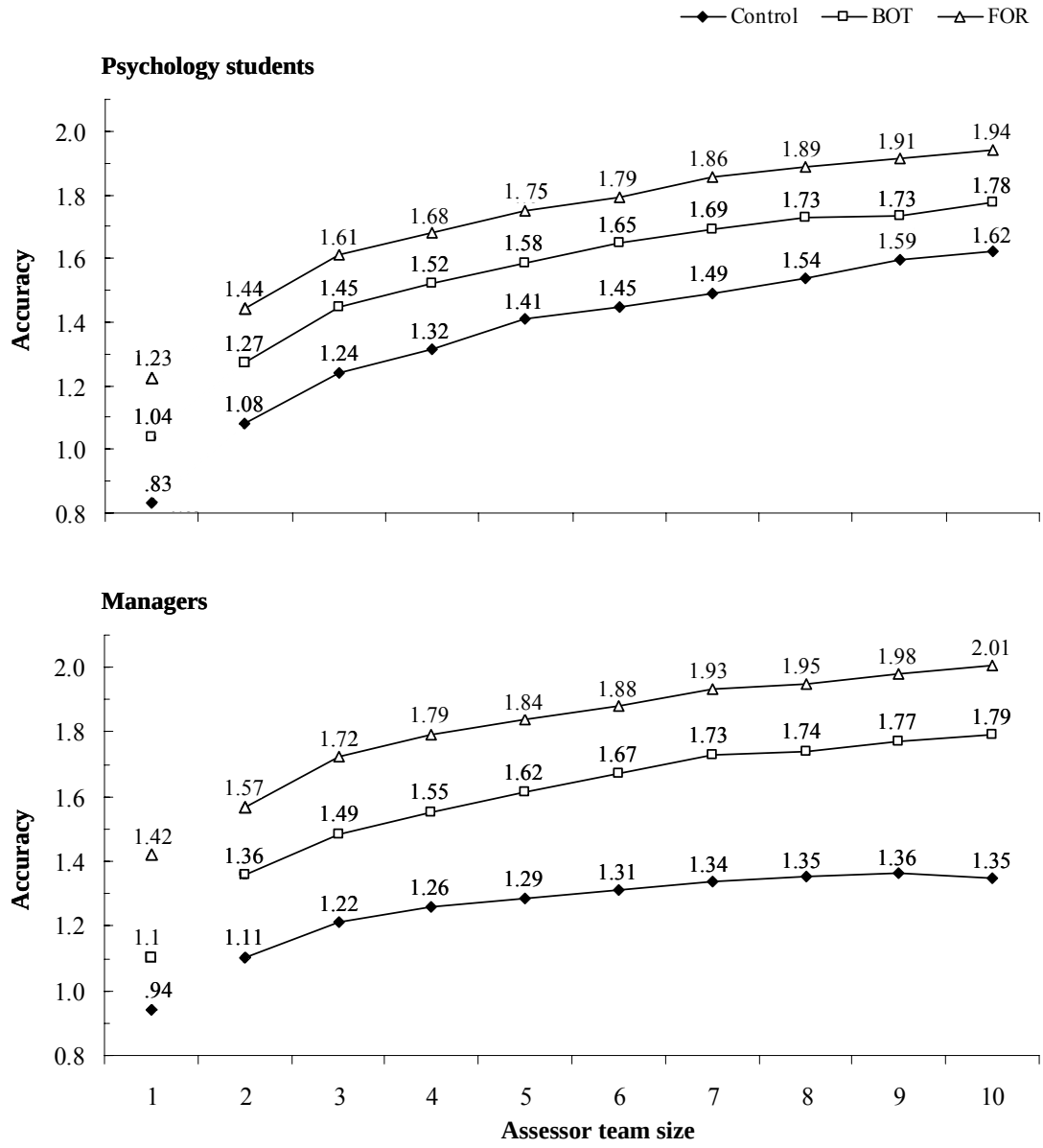

Figure 1. Average rating accuracy (Borman's differential accuracy, BDA) by assessor team size and by training condition. Higher scores indicate better accuracy. Cell-specific *n* for single psychology students (Control, *n* = 86; BOT, *n* = 73; FOR, *n* = 66) and for single managers (Control, *n* = 45; BOT, *n* = 61; FOR, *n* = 52). *n* = 1000 for all cells with a team size of ≥ 2.

Surprisingly, the main effect for assessor background was not significant, $F < 1$, indicating that in general there was no difference in rating accuracy between advanced psychology students and managers. However, this result does not mean that assessor background did not influence rating accuracy because all interaction effects involving assessor background were significant. Specifically, the interaction between assessor background and assessor training was significant and had a moderate effect on BDA, $F(2, 53946) = 4216.62$, $p < .01$, $\eta^2 = .08$, indicating that the size of the assessor training effect differed between managers and psychology students. In addition, the interaction between assessor background and assessor team size had a large effect on BDA, $F(2, 53946) = 1853.47$, $p < .01$, $\eta^2 = .15$, indicating that the effect of increasing the size of the assessor team differed between managers and psychology students. Furthermore, both the interaction between assessor team size and assessor training, $F(16, 53946) = 73.72$, $p < .01$, $\eta^2 = .01$, and the three-way interaction between assessor team size, assessor training, and assessor background, $F(16, 53946) = 73.70$, $p < .01$, $\eta^2 = .01$, had small but significant effects on BDA.

To further explore the source of the interaction effects between the investigated factors, we conducted additional analyses. One-way ANOVAs with assessor training as the independent variable revealed that the training effect was larger for managers than for psychology students (see Table 1). Thus, managers benefited more from the assessor training than psychology students. Furthermore, the effect for assessor training became more pronounced with a larger assessor team. This means that the difference in rating accuracy between untrained and trained assessors increased with an increasing size of the assessor team, especially for managers (also see Figure 1).

Table 1

Results From one-way ANOVAs With Assessor Training as the Independent Variable for Each Combination of Assessor Team Size and Assessor Background

Assessor team size	Psychology students		Managers	
	$F(2, 2997)$	η^2	$F(2, 2997)$	η^2
2	141.47**	.086	218.07**	.127
3	146.03**	.089	283.62**	.159
4	229.22**	.133	533.13**	.262
5	197.41**	.116	674.32**	.310
6	256.35**	.146	725.80**	.326
7	270.34**	.153	876.44**	.369
8	293.23**	.164	1049.50**	.412
9	239.34**	.138	1157.36**	.436
10	251.15**	.144	1491.10**	.499

Note. ** $p < .01$.

Concerning assessor team size, a larger team size was associated with a higher rating accuracy in general. However, there was one noteworthy exception from this general pattern. Specifically, untrained managers soon reached asymptotic values of rating accuracy and then did not show additional improvements of rating accuracy with an increasing size of the assessor team. In line with this, the results of one-way ANOVAs with assessor team size as the independent variable showed that the effect related to assessor team size was not even half as large for managers in the control condition as it was in any of the other cells (see Table 2), indicating that for untrained managers the effect of increasing the assessor team size was

limited. Furthermore, in contrast to untrained managers, rating accuracy for untrained psychology students improved continuously with increasing size of the assessor team. Therefore, larger teams of untrained psychology students provided more accurate ratings than untrained managers even though single managers were more accurate than single students (see Figure 2). This resulted in a significant interaction between assessor team size and assessor background when we conducted an Assessor team size × Assessor background ANOVA in the control condition, $F(8, 17982) = 33.47$, $p < .01$, $\eta^2 = .01$.

Table 2

Results From one-way ANOVAs With Assessor Team Size as the Independent Variable for Each Combination of Assessor Background and Assessor Training

	FOR		BOT		Control	
Assessor background	$F(8, 8991)$	η^2	$F(8, 8991)$	η^2	$F(8, 8991)$	η^2
Psychology students	217.64**	.162	169.10**	.131	193.00**	.147
Managers	146.60**	.115	154.30**	.121	57.49**	.049

Note. ** $p < .01$.

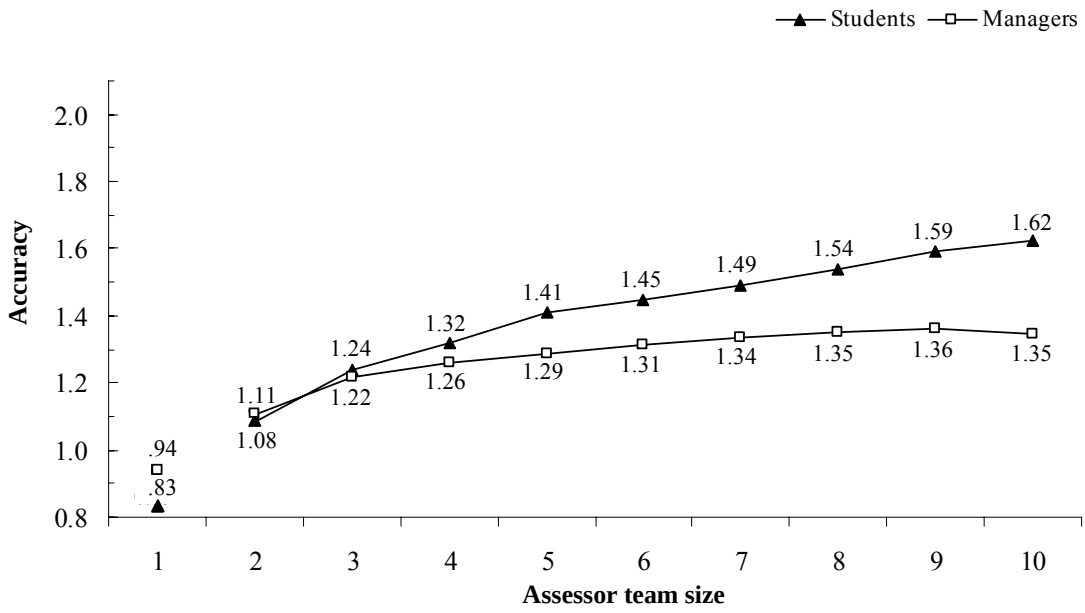

Figure 2. Average rating accuracy (Borman's differential accuracy, BDA) by assessor team size and by assessor background. Higher scores indicate better accuracy.
Cell-specific *n* for single assessors (Students, *n* = 86; Managers, *n* = 45). *n* = 1000 for all cells with a team size of ≥ 2.

Examination of Trade-Offs

First, we evaluated whether increasing the size of the assessor team can compensate for missing assessor training and vice versa. Concerning psychology students, increasing the size of the assessor team was a means to compensate for missing BOT and FOR training. For example, on average, two untrained students reached the accuracy level of a single student with BOT and three untrained students reached the accuracy level of a single student with FOR (see Figure 1). In contrast to psychology students, increasing the size of the assessor team consisting of managers could only partly compensate for missing BOT, and it was not a suitable means to compensate for a lack of FOR training. Specifically, to reach the average accuracy level of a single manager with BOT, two untrained managers sufficed. However,

untrained managers were not able to reach the accuracy level of a single FOR trained manager, even when ratings were aggregated within teams of ten assessors. Similarly, large numbers of untrained managers were also not able to outperform two managers with BOT with regard to rating accuracy.

Second, concerning assessor background, increasing the size of the assessor team could compensate for using assessors with a suboptimal background (i.e., psychology students) in all training conditions. Specifically, in all three training conditions, ratings from two and three psychology students were at least as accurate as ratings from one and two managers, respectively. In the BOT and FOR training conditions, managers were more accurate than psychology students, irrespective of the size of the assessor team, but the difference in accuracy between trained psychology students and trained managers decreased continuously with an increasing assessor team size. Conversely, with an increasing assessor team size, relations between rating accuracy of psychology students and managers changed in the control condition as already noted above. Thus, even though single untrained managers were more accurate than single untrained psychology students, untrained psychology students provided more accurate ratings than untrained managers in the case of teams of three or more assessors.

Discussion

The present study examined the effects of two factors associated with assessor expertise (assessor training and assessor background) and assessor team size on rating accuracy in an AC exercise. Of the three factors, assessor training had the largest main effect on rating accuracy. In line with psychometric theory, assessor team size also had a significant effect on rating accuracy, indicating that rating accuracy improved when ratings were aggregated across multiple assessors. However, our results suggest that increasing the size of

the assessor team can only partially compensate for missing assessor training, in particular when managers serve as assessors. Thus, appropriate assessor training seems to be essential for rating accuracy in ACs because it cannot always be substituted by aggregating ratings from multiple assessors.

An important finding of the present study is that untrained managers improved rating accuracy only to a limited degree with an increasing size of the assessor team so that even teams of ten untrained managers were unable to reach the same level of rating accuracy as a single FOR trained manager. A possible explanation for the limited effect of increasing the number of untrained managers is that single untrained managers had difficulty in differentiating between dimensions (see also Lievens, 2001a; 2001b) and rated candidates holistically instead. The present results suggest that inaccuracies of ratings due to such a holistic rating approach can be reduced only to a limited degree by aggregating multiple ratings. However, as compared to increasing the number of untrained managers, assessor training – especially in the form of FOR training – seems to be an effective means to overcome a holistic rating approach and to improve managers' rating accuracy.

In contrast to untrained managers, rating accuracy of untrained psychology students continuously improved with an increasing size of the assessor team. However, given that assessor training had the largest main effect of the three independent variables on rating accuracy, large teams of untrained psychology students were needed to reach the level of rating accuracy of a smaller team of trained psychology students, especially when the latter had taken part in FOR training. For example, to reach the rating accuracy of one, two, or three FOR trained psychology students, three, six, or ten untrained psychology students were needed, respectively. Thus, with respect to AC costs, increasing the size of the assessor team as a means to compensate for a lack of FOR training might triple personnel costs for assessors. Furthermore, it is unrealistic to assume that more than three or four assessors are

used in operational ACs to evaluate a candidate's performance in an exercise (cf. Arthur & Day, 2011; Krause & Thornton, 2009) – be it because of increased AC costs or because of decreased feasibility with an increasing number of assessors.

Finally, our results suggest that increasing the size of the assessor team can compensate for missing expertise related to assessor background. More specifically, just two and three psychology students reached the level of rating accuracy of one and two managers, respectively. Thus, using a somewhat larger number of assessors with a suboptimal background might indeed be a viable way to ensure rating accuracy, for example, under conditions when not enough assessors with sufficient expertise are available. At the same time, such moderate increases in the assessor team size might also help to keep AC costs under control.

Practical Implications

The present study provides at least three pieces of advice to users and designers of ACs. First, although assessor training is associated with higher costs for ACs, assessor training should be an inherent part of an AC program because it is an effective means to improve rating accuracy. FOR training, in which a common frame of reference for the evaluation of AC candidates is imposed on assessors, is especially recommended (see also Lievens, 2001a; Schleicher et al., 2002; Woehr & Huffcutt, 1994). When taking into account the fact that assessors can use competencies gained through assessor training each time they are employed as assessors again, the benefit of appropriate assessor training will probably outweigh training costs in the long term. Moreover, recent research has shown that beneficial effects of assessor training can also transfer to the context of performance appraisals (Macan et al., 2011) and thus go beyond the improvement of rating accuracy in ACs.

Second, if no training is provided to assessors, increasing the size of the assessor team can improve rating accuracy in an AC to some degree. However, increasing the size of the assessor team is a means to compensate for missing FOR training only if assessors have a psychological background, but not if they are managers. Yet, even if assessors have a psychological background, appropriate assessor training is probably more cost-efficient in the long-term than using larger groups of untrained assessors. Furthermore, as noted above, the effect of appropriate assessor training can transfer to other contexts such as performance appraisals (Macan et al., 2011), whereas the effect of increasing the size of the assessor team is limited to a single AC. In contrast to this, however, when conducting an AC that is only administered once or twice, increasing the size of the assessor team might be cheaper than providing extensive assessor training.

Third, the results of our study do not allow us to generally conclude whether it is more advantageous to use managers versus individuals with a psychological background as assessors in an AC. Rather, our results imply that assessors with differing backgrounds have different perspectives that might both contribute to a valuable evaluation of a candidate's performance (e.g., Damitz, Manzey, Kleinmann, & Severin, 2003). As the different perspectives due to differing assessor backgrounds "are expected and welcomed as a part of the principle of multiple assessors" (Thornton & Rupp, 2006, p. 42), we recommend using trained assessors with diverse backgrounds (see also the guidelines of the International Task Force on Assessment Center Guidelines, 2009).

Limitations and Suggestions for Future Research

Some limitations of this study should be mentioned. First, the analyses were based on data from a simulated selection situation setting, in which assessors were managers and I\O psychology students, respectively. Hence, it is unclear to what degree our results generalize to

professional psychologists (or experienced HR professionals in general) who have specialized in conducting ACs and who regularly serve as assessors in ACs. In addition, professional psychologists are trained to base personality judgments on behavioral observations and to differentiate between traits (Sagie & Magnezy, 1997), which also might be beneficial for providing accurate ratings. Therefore, and in light of previous findings (e.g., Gaugler et al., 1987; Sagie & Magnezy, 1997; Woehr & Arthur, 2003), we would expect professional psychologists to generally outperform managerial assessors with respect to their rating accuracy. Furthermore, the effect of assessor training might be less pronounced for professional psychologists than in this study because professional psychologists might hold more appropriate performance schemas in the first place because of both their background and experience.

Second, the exercise used in this study was a presentation exercise in which assessors observed only one candidate. When assessors have to observe multiple candidates simultaneously as, for example, in a group discussion, cognitive demands increase and thus inaccuracies in ratings are likely to increase, too (cf. Melchers, Kleinmann, & Prinz, 2010). Therefore, assessor expertise as well as increasing the number of assessors and aggregating their ratings might be particularly important in exercises with increased cognitive demands. Future research with regard to this issue is needed.

And third, in the present study, the nature of the stimulus materials did not allow us to determine construct-related and criterion-related validity of the dimension ratings. Therefore, we focused on rating accuracy as the dependent variable. However, previous findings suggest that factors that improve rating accuracy usually also lead to improvements in construct-related and criterion-related validity (Gaugler & Thornton, 1989; Lievens, 2001a; Melchers et al., 2010; Schleicher et al., 2002).Therefore, we assume that assessor training, assessor background, and assessor team size have similar effects on construct-related and criterion-

related validity of ACs as they have on rating accuracy. Nevertheless, future research is needed to confirm that increasing the size of the assessor team is a means to compensate for missing assessor expertise with regard to AC validity.

References

Arthur, W., Jr., & Day, E. A. (2011). Assessment centers. In S. Zedeck (Ed.), *APA handbook of industrial and organizational psychology: Selecting and developing members for the organization* (Vol. 2, pp. 205-235). Washington, DC: APA.

Arthur, W., Jr., Day, E. A., McNelly, T. L., & Edens, P. S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology, 56*, 125-154. doi:10.1111/j.1744-6570.2003.tb00146.x

Bernardin, H. J., Buckley, M. R., Tyler, C. L., & Wiese, D. S. (2000). A reconsideration of strategies for rater training. In G. R. Ferris (Ed.), *Research in Personnel and Human Resources Management* (Vol. 18, pp. 221-274). Greenwich, CT: JAI Press.

Borman, W. C. (1977). Consistency of rating accuracy and rating errors in the judgment of human performance. *Organizational Behavior and Human Performance, 20*, 238-252. doi:10.1016/0030-5073(77)90004-6

Cardy, R. L., Bernardin, H. J., Abbott, J. G., Senderak, M. P., & Taylor, K. (1987). The effects of individual performance schemata and dimension familiarization on rating accuracy. *Journal of Occupational and Organizational Psychology, 60*, 197-205. doi:10.1111/j.2044-8325.1987.tb00253.x

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2 ed.). Hillsday, NJ: Lawrence Erlbaum.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Damitz, M., Manzey, D., Kleinmann, M., & Severin, K. (2003). Assessment center for pilot selection: Construct and criterion validity and the impact of assessor type. *Applied Psychology: An International Review, 52*, 193-212. doi:10.1111/1464-0597.00131

Dilchert, S., & Ones, D. S. (2009). Assessment center dimensions: Individual differences correlates and meta-analytic incremental validity. *International Journal of Selection and Assessment, 17*, 254-270. doi:10.1111/j.1468-2389.2009.00468.x

Epstein, S. (1983). Aggregation and beyond: Some basic issues on the prediction of behavior. *Journal of Personality, 51*, 360-392. doi:10.1111/j.1467-6494.1983.tb00338.x

Eurich, T. L., Krause, D. E., Cigularov, K., & Thornton, G. C., III. (2009). Assessment centers: Current practices in the United States. *Journal of Business and Psychology, 24*, 387-407. doi:10.1007/s10869-009-9123-3

Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., III, & Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology, 72*, 493-511. doi:10.1037/0021-9010.72.3.493

Gaugler, B. B., & Thornton, G. C., III. (1989). Number of assessment center dimensions as a determinant of assessor accuracy. *Journal of Applied Psychology, 74*, 611-618. doi:10.1037/0021-9010.74.4.611

Gorman, C. A., & Rentsch, J. R. (2009). Evaluating frame-of-reference rater training effectiveness using performance schema accuracy. *Journal of Applied Psychology, 94*, 1336-1344. doi:10.1037/a0016476

Hardison, C. M., & Sackett, P. R. (2004). *Assessment center criterion related validity: A meta-analytic update*. Unpublished manuscript.

Hermelin, E., Lievens, F., & Robertson, I. T. (2007). The validity of assessment centres for the prediction of supervisory performance ratings: A meta-analysis. *International Journal of Selection and Assessment, 15*, 405-411. doi:10.1111/j.1468-2389.2007.00399.x

International Task Force on Assessment Center Guidelines. (2009). Guidelines and ethical considerations for assessment center operations. *International Journal of Selection and Assessment, 17*, 243-253. doi:10.1111/j.1468-2389.2009.00467.x

Kozlowski, S. W. J., Kirsch, M. P., & Chao, G. T. (1986). Job knowledge, ratee familiarity, conceptual similarity and halo error: An exploration. *Journal of Applied Psychology, 71*, 45-49. doi:10.1037/0021-9010.71.1.45

Krause, D. E., & Gebert, D. (2003). A comparison of assessment center practices in organizations in german-speaking regions and the United States. *International Journal of Selection and Assessment, 11*, 297-312. doi:10.1111/j.0965-075X.2003.00253.x

Krause, D. E., Kersting, M., Heggestad, E. D., & Thornton, G. C., III. (2006). Incremental validity of assessment center ratings over cognitive ability tests: A study at the executive management level. *International Journal of Selection and Assessment, 14*, 360-371. doi:10.1111/j.1468-2389.2006.00357.x

Krause, D. E., & Thornton, G. C., III. (2009). A cross-cultural look at assessment center practices: Survey results from Western Europe and North America. *Applied Psychology: An International Review, 58*, 557-585. doi:10.1111/j.1464-0597.2008.00371.x

Lievens, F. (1999). Development of a simulated assessment center. *European Journal of Psychological Assessment, 15*, 117-126. doi:10.1027//1015-5759.15.2.117

Lievens, F. (2001a). Assessor training strategies and their effects on accuracy, interrater reliability, and discriminant validity. *Journal of Applied Psychology, 86*, 255-264. doi:10.1037/0021-9010.86.2.255

Lievens, F. (2001b). Assessors and use of assessment centre dimensions: A fresh look at a troubling issue. *Journal of Organizational Behavior, 22*, 203-221. doi:10.1002/job.65

Lievens, F., & Klimoski, R. J. (2001). Understanding the assessment center process: Where are we now? In C. L. Cooper & I. T. Robertson (Eds.), *International Review of Industrial and Organizational Psychology* (Vol. 16, pp. 245-286). Chichester, UK: Wiley.

Macan, T., Mehner, K., Havill, L., Meriac, J. P., Roberts, L., & Heft, L. (2011). Two for the price of one: Assessment center training to focus on behaviors can transfer to performance appraisals. *Human Performance, 24*, 443-457. doi:10.1080/08959285.2011.614664

Melchers, K. G., & Annen, H. (2010). Officer selection for the Swiss armed forces: An evaluation of validity and fairness issues. *Swiss Journal of Psychology, 69*, 105-115. doi:10.1024/1421-0185/a000012

Melchers, K. G., Kleinmann, M., & Prinz, M. A. (2010). Do assessors have too much on their plates? Rating quality and the number of simultaneously observed candidates in assessment center group discussions. *International Journal of Selection and Assessment, 18*, 329-341. doi:10.1111/j.1468-2389.2010.00516.x

Meriac, J. P., Hoffman, B. J., Woehr, D. J., & Fleisher, M. S. (2008). Further evidence for the validity of assessment center dimensions: A meta-analysis of the incremental criterion-related validity of dimension ratings. *Journal of Applied Psychology, 93*, 1042-1052. doi:10.1037/0021-9010.93.5.1042

Noe, R. A. (2002). *Employee training and development* (2. ed.). New York: McGraw-Hill.

Sagie, A., & Magnezy, R. (1997). Assessor type, number of distinguishable dimension categories, and assessment centre construct validity. *Journal of Occupational and Organizational Psychology, 70*, 103-108. doi:10.1111/j.2044-8325.1997.tb00634.x

Schleicher, D. J., & Day, D. V. (1998). A cognitive evaluation of frame-of-reference rater training: Content and process issues. *Organizational Behavior and Human Decision Processes, 73*, 76-101. doi:10.1006/obhd.1998.2751

Schleicher, D. J., Day, D. V., Mayes, B. T., & Riggio, R. E. (2002). A new frame for frame-of-reference training: Enhancing the construct validity of assessment centers. *Journal of Applied Psychology, 87*, 735-746. doi:10.1037/0021-9010.87.4.735

Sulsky, L. M., & Balzer, W. K. (1988). Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. *Journal of Applied Psychology, 73*, 497-506. doi:10.1037/0021-9010.73.3.497

Thornton, G. C., & Rupp, D. E. (2006). *Assessment centers in human resource management: Strategies for prediction, diagnosis, and development*. Mahwah, NJ: Lawrence Erlbaum.

Woehr, D. J., & Arthur, W., Jr. (2003). The construct-related validity of assessment center ratings: A review and meta-analysis of the role of methodological factors. *Journal of Management, 29*, 231-258. doi:10.1177/014920630302900206

Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology, 67*, 189-205. doi:10.1111/j.2044-8325.1994.tb00562.x

Chapter 2

Are Improvements in Assessment Center Construct-Related Validity Paralleled by Improvements in Criterion-Related Validity? The Effects of Exercise Similarity on Assessment Center Validity

Andreja Wirz, Klaus G. Melchers, Stefan Schultheiss, and Martin Kleinmann

Universität Zürich

Acknowledgements

We thank Sabrina Engeli and Pascale Gschwend for their help with data collection.

Manuscript submitted for publication to *Human Performance*

Abstract

We followed recent calls to evaluate construct-related and criterion-related validity of assessment centers (ACs) simultaneously. Specifically, we examined the effects of exercise similarity on both aspects of validity within a single study. Exercise similarity was operationalized by using two different types of exercises. Data were collected in an AC that consisted of presentation exercises and leaderless group discussions ($N = 92$). As expected, convergent validity was better for similar exercises than it was for dissimilar exercises. However, regarding criterion-related validity, we did not find differences between similar and dissimilar exercises. Hence, this study revealed that improvements in construct-related validity are not necessarily paralleled by improvements in criterion-related validity. Practical implications and directions for future research are discussed.

According to the unitarian framework of validity (e.g., Binning & Barrett, 1989; Landy, 1986; Messick, 1995), the construct-related and criterion-related validity of a test are closely connected to each other, so that improvements in construct-related validity are assumed to lead to improvements in criterion-related validity and vice versa. The unitarian framework of validity should be applicable to all instruments for selection and assessment. However, in the assessment center (AC) domain, research related to construct-related and criterion-related validity has largely evolved independently, making it difficult to allow clear statements about how construct-related and criterion-related validity of ACs are related to each other. Because of this, several authors recently called for a broad validation strategy and to examine both aspects of validity simultaneously (e.g., Lievens, 2009; Lievens, Dilchert, & Ones, 2009; Melchers & König, 2008; Woehr & Arthur, 2003). This call is of particular importance given recent arguments to abandon research concerning possible means to improve the construct-related validity of ACs (Lance, 2008). Such a step seems premature given the limited available knowledge of how moderators of construct-related validity influence criterion-related validity (Melchers & König, 2008).

In line with the unitarian framework of validity, the few studies that are available to date revealed that improvements in AC construct-related validity may indeed lead to improvements in criterion-related validity (Melchers, Kleinmann, & Prinz, 2010; Schleicher, Day, Mayes, & Riggio, 2002). Other findings, however, suggest that some factors might have opposite effects on construct-related and criterion-related validity of ACs. Specifically, using a set of similar exercises is beneficial for construct-related validity (Highhouse & Harris, 1993; Sackett & Harris, 1988; Schneider & Schmitt, 1992), but it might not be advisable for criterion-related validity (Gaugler, Rosenthal, Thornton, & Bentson, 1987; see also Lievens et al., 2009). However, findings concerning the effects of exercise similarity on construct-related and criterion-related validity of ACs stem from independent studies and thus do not allow a

definite conclusion about whether exercise similarity indeed has diverging effects on construct-related and criterion-related validity of ACs. Therefore, we aim to examine the effects of exercise similarity on both aspects of validity within a single study. In doing so, we expand the sparse research on the effects that factors simultaneously have on construct-related and criterion-related validity of ACs. Thereby, our findings will contribute to a greater understanding of the connection between the construct-related and criterion-related validity of ACs that is relevant particularly when interventions to improve one aspect of validity are implemented.

The Unitarian Framework of Validity

Proponents of the unitarian framework of validity (Binning & Barrett, 1989; Landy, 1986; Messick, 1995) have argued that content-related, construct-related, and criterion-related validity of a test build a logically linked system and thus can be unified within a concept. Specifically, Binning and Barrett stated that support for two of the three aspects of validity evidenced that the third aspect was present, too. More precisely, Binning and Barrett conclude that "if it can be shown that a test measures a specific construct that has been determined to be critical for job performance, then inferences about job performance from test scores are, by logical implication, justified" (p. 482). In line with this, improvements in construct-related validity are assumed to be paralleled by improvements in criterion-related validity and vice versa.

The unitarian framework of validity should apply to all instruments for personnel selection and assessment and thus also to ACs. Therefore, we will review research on the content-related, criterion-related, and construct-related validity of ACs in the next section to show to which degree the assumptions of the unitarian framework of validity have been supported in the AC domain so far.

Assessment Center Validity

Content-related validity refers to the job-relatedness of the AC that is established when exercises represent situations that are critical to the target job and when they challenge candidates with the type of problems that occur in the target job (Thornton & Rupp, 2006). In addition, Arthur and Day (2010) pointed out that a proper dimension explication is a further component of content-related validity. Furthermore, the appropriateness of instructions to candidates and of the scoring system also contribute to the job-relatedness of an AC (Sackett, 1987). Job analysis is viewed as a cornerstone for establishing job-relatedness of ACs and is, therefore, defined as an essential element of ACs (International Task Force on Assessment Center Guidelines, 2009). Most organizations conduct a job analysis prior to designing an AC (Eurich, Krause, Cigularov, & Thornton, 2009; Krause & Gebert, 2003; Krause & Thornton, 2009) and thus ACs are considered to have content-related validity (cf. Woehr & Arthur, 2003).

Concerning criterion-related validity, research has repeatedly demonstrated that ACs are criterion valid. Meta-analytically estimated criterion-related validities for the overall AC rating range from .26 to .40, indicating that the overall AC rating allows good predictions of job performance (Becker, Höft, Holzenkamp, & Spinath, 2011; Gaugler et al., 1987; Hardison & Sackett, 2004; Hermelin, Lievens, & Robertson, 2007). Additional meta-analytic findings also suggest that AC scores are criterion valid on the dimension-level and that some dimensions (e.g., organizing and planning) are more predictive of job performance than others (e.g., consideration of others; Arthur, Day, McNelly, & Edens, 2003; Meriac, Hoffman, Woehr, & Fleisher, 2008). Moreover, AC ratings have incremental validity for the prediction of job performance even beyond that of cognitive ability tests and personality inventories (Dilchert & Ones, 2009; Krause, Kersting, Heggestad, & Thornton, 2006; Melchers & Annen, 2010; Meriac et al., 2008).

Despite the evidence for content-related and criterion-related validity of ACs, findings concerning the construct-related validity of ACs are less promising and so it is controversial as to what degree ACs measure the purported dimensions. Usually, correlations between ratings on a specific dimension from different exercises are low and thus convergent validity is poor (cf. Melchers, Henggeler, & Kleinmann, 2007; and Woehr & Arthur, 2003, for meta-analytic results). Furthermore, dimension ratings within exercises usually correlate substantially, indicating that they lack discriminant validity. In addition, confirmatory factor analyses have revealed a similar picture. If dimension factors could be found at all, they usually explained a smaller proportion of variance in AC ratings than exercise factors (e.g., Bowler & Woehr, 2006; Hoffman, Melchers, Blair, Kleinmann, & Ladd, 2011; Lance, Lambert, Gewin, Lievens, & Conway, 2004; Lievens et al., 2009).

Based on the findings above, one might conclude that ACs are criterion valid (as well as content valid) while having problems with regard to their internal structure, which seems to be inconsistent with the unitarian framework of validity (e.g., Binning & Barrett, 1989). However, even though large bodies of research are available concerning both construct-related as well as criterion-related validity, research related to each aspect of AC validity has evolved in a largely unconnected fashion, that is, most of the studies focused either on the construct-related or on the criterion-related validity of ACs. Thus, it remains unclear whether ACs are criterion valid and at the same time really lacking construct-related validity and whether improvements in one aspect of validity are paralleled by improvements in the other aspect. Because of this, several authors have called for future research to follow a broader validation strategy and to investigate both aspects of validity simultaneously (e.g., Lievens, 2009; Lievens et al., 2009; Melchers & König, 2008; Woehr & Arthur, 2003).

To date, only a handful of studies have investigated both construct-related and criterion-related validity of an AC simultaneously (Chan, 1996; Fleenor, 1996; Henderson,

Anderson, & Rick, 1995; Jansen & Stoop, 2001; Lievens et al., 2009; see also Woehr & Arthur, 2003). Except for one of these studies (Chan, 1996), all revealed results that were consistent with the unitarian framework of validity. For example, Henderson et al. (1995) found correlations between AC scores and job-related criteria to be low when evidence of construct-related validity was weak. Furthermore, Lievens et al. (2009) found that dimensions that explained more variance in AC ratings were also more predictive of job-related criteria compared to dimensions that explained less variance in AC ratings. Lievens et al. concluded from their results that "evidence of internal construct-related validity appears to be coupled with evidence of criterion-related validity" (p. 386). These findings suggest that construct-related and criterion-related validity of ACs are connected to each other as assumed by the unitarian framework of validity.

However, these results do not lend an answer to the question of whether improvements in one aspect of validity are paralleled by improvements in the other aspect. So far, only two studies have examined the effects of variations of AC design factors on construct-related and criterion-related validity simultaneously. In line with the unitarian framework of validity, they found parallel effects on both aspects of validity. Schleicher et al. (2002) found that when assessors took part in frame-of-reference rater training, construct-related as well as criterion-related validity were better compared to when assessors were untrained. Because assessors provided more reliable and more accurate ratings after frame-of-reference training than after control training, the improvement in both construct-related and criterion-related validity after frame-of-reference training can be attributed to an improvement in the reliability and accuracy of dimension ratings. Furthermore, in a study by Melchers et al. (2010) that focused on AC group exercises, both aspects of validity improved when assessors had to observe a smaller compared to a larger number of candidates simultaneously. In addition to this, meta-analytic findings provide indirect support that specific factors have parallel effects on construct-related

and criterion-related validity. For example, ACs have been found to be more construct valid (Woehr & Arthur, 2003) and more criterion valid (Gaugler et al., 1987) when psychologists served as assessors compared to when managers served as assessors.

Effects of Exercise Similarity on Assessment Center Validity

Despite the initial findings from Melchers et al. (2010) and Schleicher et al. (2002), it might be premature to conclude that improvements in construct-related validity of ACs are always paralleled by improvements in criterion-related validity. Instead, some factors might have differing effects on construct-related and on criterion-related validity. Specifically, exercise similarity might improve construct-related validity but potentially lead to impairments in criterion-related validity.

Effects of exercise similarity on construct-related validity. In discussions on potential causes for the lack of construct-related validity of AC ratings, several authors emphasized the characteristics of AC exercises and the consequences of these characteristics for behavioral consistency. Usually, ACs are designed to represent a broad range of contextual demands of the target position and thus are comprised of a diverse set of exercises. Thus, Neidig and Neidig (1984) argued that different exercises require different kinds of behaviors. Other authors stated that exercises differ in the opportunity to manifest behavior that is related to a particular dimension (Sackett & Dreher, 1982) and that exercises elicited different facets of a particular dimension (Howard, 2008). As a result of this, the candidates' performance may differ across different exercises and thus low convergence between dimension ratings across exercises is not surprising (e.g., Highhouse & Harris, 1993). However, when using a set of exercises with similar characteristics, that is, when exercises pose similar demands on candidates, correlations between ratings on identical dimensions can be expected to increase (e.g., Neidig & Neidig, 1984; see also Sackett & Harris, 1988).

Several studies found support for the assumption that exercise similarity is related to the construct-related validity of AC ratings. For example, Sackett and Harris (1988) showed that dimension factors are more likely to be found when exercises are structurally similar. Specifically, evidence for dimension factors was substantial for an AC that consisted of group discussions only, but not for ACs that were comprised of structurally different exercises, for example, group discussions, in-basket exercises, and role plays. Schneider and Schmitt (1992) directly examined the effects of exercise type and exercise content on the variance of AC dimension ratings. While the effect of exercise content (i.e., of using exercises that required either cooperative or competitive behavior) on the convergence of dimension ratings was negligible, exercise type accounted for a substantial amount of variance in dimension ratings, indicating that ratings on identical dimensions converged more across exercises of the same type (e.g., across two group discussions) than across exercises of different types (e.g., across a group discussion and a role play). Finally, Highhouse and Harris (1993) found that convergence between dimension ratings was better across exercises that were perceived as similar in terms of behavioral requirements. Taken together, exercise similarity seems to be beneficial for construct-related validity of dimension ratings, in particular for convergent validity.

A theoretical explanation for the effect of exercise similarity on construct-related validity of ACs is offered by trait activation theory (TAT, Tett & Burnett, 2003; Tett & Guterman, 2000). According to TAT, situations differ in the degree to which they provide trait-relevant cues, which means that situations differ in their potential to elicit behavior that is related to a specific trait. Behavioral consistency across situations can only be expected when situations are similar with regard to their trait-relevance and if the situations' potential to activate specific traits is high. In line with this, studies that applied TAT in the domain of ACs found that ratings of dimensions linked to a given Big Five trait showed stronger

convergence when they stemmed from exercises judged to be high in trait-activation potential for this trait than ratings from exercises that were low in trait-activation potential for this trait (Haaland & Christiansen, 2002; Lievens, Chasteen, Day, & Christiansen, 2006). With regard to exercise similarity, TAT suggests that similar exercises will more likely activate comparable traits than dissimilar exercises would. This in turn would be beneficial for the convergent validity of ratings of dimensions linked to the relevant trait.

Effects of exercise similarity on criterion-related validity. While using similar exercises instead of dissimilar exercises seems beneficial for construct-related validity, criterion-related validity might decrease when the range of exercises is limited (Lievens et al., 2009) because increased similarity of the exercises that is associated with behavioral consistency might restrict the range of observable behaviors. A set of different exercises should elicit a broader range of job-related behaviors than a comparable set of similar exercises (cf. Neidig & Neidig, 1984) and the larger the number of job-related behaviors that can be observed, the better the AC's potential to predict candidates' job performance. Since a set of diverse exercises potentially samples the situational demands of a target job more comprehensively than a set of similar exercises, exercise diversity should be beneficial for AC criterion-related validity. Conversely, high exercise similarity might limit criterion-related validity so that criterion-related validity might be better when exercises are dissimilar compared to when they are similar. In line with this, Gaugler et al. (1987) presented meta-analytic evidence that using a larger number of *different* exercises in ACs is associated with better criterion-related validity.

Limitations of Previous Research

Taken together, the arguments and findings described above suggest that exercise similarity might lead to an improvement in construct-related validity of ACs, while criterion-

related validity of ACs might decrease when using similar instead of dissimilar exercises. However, findings on the effects of exercise similarity on construct-related and on criterion-related validity of ACs stem from independent studies that always focused on only one aspect of validity. Therefore, we aimed to clarify the effects of exercise similarity on construct-related and criterion-related validity within a single study. In line with previous research, we expected the convergent validity of AC dimension ratings to be better when exercises are similar compared to when they are dissimilar. With regard to criterion-related validity we intended to investigate whether criterion-related validity of AC dimension ratings is indeed better when exercises are dissimilar compared to when they are similar or whether the expected improvements of convergent validity when using similar exercises are paralleled by improvements in criterion-related validity.

Method

For the purpose of the present study, we conducted a simulated one-day graduate AC. Based on findings from Schneider and Schmitt (1992), we operationalized exercise similarity through exercise type. More precisely, we used presentations and group discussions as dissimilar exercises and conducted an AC that consisted of these two types of exercises.

Recent and prospective university graduates were invited via e-mail to participate in the AC with the opportunity to gain experience in selection situations and to receive feedback pertaining to their performance. Conditions for participation were that participants were employed more than 12 hours per week during a six month period before the AC and that they agreed that their supervisors were asked to evaluate their job performance.

Sample

A total of 117 participants took part in the AC. Twenty-five participants were excluded from the analyses either because they did not complete all AC exercises, no criterion data were available for them, or because their supervisors reported difficulties when providing criterion data. Thus, data from 92 participants (50% males, 50% females) could be used for the analyses. The participants' age ranged from 22 to 58 years (M = 29.10, SD = 6.20). Most of the participants held a Bachelor's degree (22.8%) or a Master's degree (47.8%), mainly in natural sciences (25.0%), social sciences (22.8%), or business and economics (18.5%). Almost half of the participants (46.7 %) were working in education and research, and nearly 10% each in banking and insurance and in the service industry. The participants' average job tenure ranged from 3 months to 15 years and was 2.83 years on average.

Assessment Center Design

The AC was designed to simulate a one-day graduate assessment because graduate trainee positions cover a wide range of requirements that are essential in many jobs, such as, for example, analyzing documents, organizing and presenting information, and working out solutions in groups. The AC consisted of presentations and leaderless group discussions (LGDs). One presentation exercise (Presentation 1) was a sales presentation in which participants had to persuade a potential client of a fictitious company to purchase a manufacturing system. In the other presentation exercise (Presentation 2), participants were asked to present a leisure activity of their own choice to a group of other job starters. Examples of chosen leisure activities were sports like volleyball, snowboarding, and hiking, cultural interests like literature, photography, and playing the flute, and others like, for example, financial markets.

One of the group discussions was a staffing task (LGD 1). The group had to identify the best applicant for a vacant position in a fictitious credit bank. To find the proper solution, participants needed to collaborate and to share previously received information on the applicants that was distributed among the group. In the second group discussion (LGD 2), participants first had to individually rank ten graduate marketing activities according to their perceived efficacy. The group then had to discuss the graduate marketing activities and to find a common rank order. Participants were instructed that the common rank order should correspond as much as possible with their individual rank order. In the third group discussion (LGD 3), participants received the same instructions as in the previous group discussion, but they had to discuss ten activities for improving the work-life-balance of a company's employees. For each exercise, participants had 15 to 20 minutes preparation time. The presentations lasted 10 minutes and each group discussion lasted 30 minutes.

In the AC, participants were evaluated on six dimensions, namely analytical skills, persuasiveness, organizing and planning, assertiveness, cooperation, and presentation skills. Dimension definitions can be found in Table 1, and Table 2 shows the dimension by exercise matrix. The dimension definitions, the exercise instructions, and the dimension by exercise matrix were reviewed by five subject matter experts who were familiar with assessment center research and practice and who discussed the dimension definitions and the appropriate exercise by dimension matrix until consensus was reached.

Table 1

Dimension Definitions

Dimension	Definition
Analytical skills	Analyzing carefully; quickly and correctly comprehending new contents; correctly recognizing connections; differentiating between important and unimportant.
Persuasiveness	Clearly explaining one's decisions; presenting solid arguments; selling one's ideas to others.
Organizing and planning	Being systematic; differentially organizing information; structuring presentations or discussions in a useful way; adequately estimating time requirements.
Assertiveness	Pushing one's interests even in light of resistance from others; not letting oneself get discouraged by others; acting in a determined way.
Cooperation	Picking up ideas that differ from one's own view; being willing to adapt one's view; helping to achieve objectives of the group.
Presentation skills	Appearing confident; speaking calmly and clearly; using gestures and mimic to support the verbal; turning toward listeners; maintaining eye contact with listeners.

Table 2

Dimension by Exercise Matrix

Dimension	Presentation 1	Presentation 2	LGD 1	LGD 2	LGD 3
Analytical skills	X		X		
Persuasiveness	X	X	X	X	X
Organizing and planning	X	X	X	X	X
Assertiveness				X	X
Cooperation			X		
Presentation skills	X	X			

Note. LGD = leaderless group discussion. Presentation 1 = sales presentation, Presentation 2 = leisure activity presentation, LGD 1 = staffing task, LGD 2 = graduate marketing task, LGD 3 = work-life-balance task.

Procedure

Six to twelve participants took part in each AC. At the beginning, participants received general information on the AC method. Furthermore, participants were told to imagine they were candidates in a graduate AC of a telecommunications company. As an incentive to strive for good performance, the best candidate in each administration of the AC was promised to receive 100 Swiss francs, and the second best candidate about 50 Swiss francs.

In each exercise, two assessors observed a candidate and independently rated his or her performance on three to four pre-defined dimensions (see Table 2). After the completion of all exercises, assessors who observed the same candidate in an exercise met for discussion. If ratings on a dimension diverged by more than one point, assessors had to discuss and adjust

their ratings. Afterwards, assessors gave feedback to participants and rewarded the best and second best candidate of the day.

Assessors

Assessors were Master's level psychology students (11 males, 23 females) with an average age of 26.85 years (*SD* = 2.34). Prior to the AC, assessors took part in a one-day rater training that was a combination of behavior observation training and frame-of-reference training (cf. Woehr & Huffcutt, 1994). At the beginning of the rater training, assessors received general information on ACs. Then, they were familiarized with the exercises, were introduced to the dimension definitions, and trainers presented examples of behaviors for good and poor performance on each dimension, respectively. Next, assessors practiced distinguishing between observing and evaluating a candidate's performance. Assessors also received frame-of-reference training to establish a common evaluation standard. To this end, assessors observed and evaluated the performance of one to three other assessors who simulated an AC exercise. Afterwards, assessors and trainers discussed the performance evaluations and clarified discrepancies among ratings. Finally, assessors received instructions for the discussion and the feedback procedure. Eight assessors were not able to participate in the rater training. Therefore, seven of them shadowed a trained assessor in an AC and one of them was individually trained with videotaped performances before they served as full assessors whose ratings were used for the study.

Exercise Similarity

As mentioned above, presentation exercises and group discussions were used in the present study. Exercises of the same type were considered to be similar to each other (cf. Schneider & Schmitt, 1992). Accordingly, presentation exercises and group discussions were regarded as dissimilar. To test whether the similarity classification according to exercise type

was justified, experienced assessors evaluated the similarity of each pair of exercises on a 7-point scale (ranging from 1 = *not similar at all* to 7 = *absolutely similar*). These experienced assessors were ten raters who had served at least four times as assessors in the AC and thus were familiar with the exercises. On average, they rated exercise pairs of the same type as being more similar ($M = 4.38$, $SD = 1.30$) than exercise pairs of different types ($M = 1.90$, $SD = 0.96$), $t(9) = 6.88$, $p < .01$. Hence, perceived exercise similarity supported the categorization of similar and dissimilar exercises according to exercise type.

Measures

Assessment center performance. In each exercise, assessors evaluated the participants' performance on the pre-defined dimensions on a five-point scale (from 1 = *poor* to 5 = *excellent*). To determine interrater reliability, we calculated intraclass correlations (ICC 1.1) between the post-discussion dimension ratings of the two assessors who evaluated a candidate in an exercise. The average intraclass correlation of the post-discussion dimension ratings and thus the reliability of a single assessor was $r = .72$. We calculated average post-consensus dimension ratings from the two assessors, mean dimension ratings across exercises, and an overall AC rating that represented the statistical mean across all exercises and dimensions.

Job performance. Shortly before the AC, the participants' supervisors were asked to complete an online questionnaire to evaluate the participants' job performance. Supervisors completed five items from the task-based job performance questionnaire by Bott, Svyantek, Goodman, and Bernal (2003) and five items from the German translation (Staufenbiel & Hartz, 2000) of Williams' and Anderson's (1991) in-role behavior scale.

Goffin, Jelley, Powell, and Johnston (2009) have demonstrated that performance ratings are more valid when a rating method is used that encourages social comparisons

instead of a non-comparative method. Therefore, we instructed supervisors to evaluate the participants' performance in comparison to the participants' colleagues or in comparison to former employees in a similar position. In addition, we adapted the formulation of the items so that they included a social comparison. Examples of the items used are "In comparison to his or her colleagues, he [or she] demonstrates expertise in all job related tasks" and "In comparison to his [or her] colleagues, he [or she] meets formal performance requirements of the job". Ratings were made on a 7-point scale (ranging from 1 = *not at all* to 7 = *absolutely*) The job performance score used for the analyses was the average score across all ten items, the coefficient alpha of which was .92.

Results

Preliminary Analyses

To determine the realism of our AC, we asked participants about their behavior in the AC simulation. Of the participants, 90.20 % indicated that they acted like they would in a real selection situation.

Correlations between participants' demographic variables, AC performance, and job performance are reported in Table 3. Correlations indicate that men generally performed better in the AC than women. Age was related to neither AC performance nor to evaluations of job performance. The correlation between the overall AC rating and job performance was significant, $r = .21$, $p < .05$, and comparable to meta-analytic estimates of criterion-related validity (see Gaugler et al., 1987; Hardison & Sackett, 2004; Hermelin et al., 2007). On the dimension-level, criterion-related validity coefficients ranged between $r = .07$, ns, (for assertiveness and cooperation) and $r = .29$, $p < .01$, (for organizing and planning).

Table 3

Means, Standard Deviations, and Correlations Between Candidates' Demographic Variables, AC Performance, and Job Performance

Variable	*M*	*SD*	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.
1. Gender	1.50	0.50															
2. Age	29.10	6.19	.09														
3. Presentation 1	3.48	0.86	-.40**	.07	*(.88)*												
4. Presentation 2	3.47	0.74	-.28**	.13	.50**	*(.72)*											
5. LGD 1	2.95	0.73	-.26*	-.06	.36**	.38**	*(.81)*										
6. LGD 2	3.21	0.84	-.18	.16	.50**	.44**	.49**	*(.80)*									
7. LGD 3	3.18	0.74	-.14	.10	.33**	.40**	.60**	.57**	*(.70)*								
8. Analytical Skills	3.17	0.72	-.35**	-.12	.75**	.31**	.63**	.48**	.44**	*(.35)*							
9. Persuasiveness	3.41	0.59	-.34**	.09	.69**	.65**	.64**	.76**	.67**	.65**	*(.69)*						
10. Organizing and planning	3.19	0.73	-.29**	.11	.66**	.71**	.71**	.70**	.68**	.56**	.74**	*(.76)*					
11. Assertiveness	3.00	0.88	-.14	.16	.35**	.41**	.50**	.82**	.80**	.42**	.70**	.60**	*(.66)*				
12. Cooperation	3.14	0.89	-.19	-.09	.19	.24*	.78**	.20	.45**	.34**	.35**	.47**	.29**	-			
13. Presentation skills	3.42	0.84	-.31**	.18	.75**	.70**	.37**	.45**	.32**	.47**	.57**	.63**	.31**	.19	*(.71)*		
14. Overall AC rating	3.26	0.59	-.34**	.11	.73**	.72**	.74**	.81**	.76**	.70**	.91**	.92**	.76**	.48**	.70**	*(.81)*	
15. Job performance	5.85	0.86	.02	-.01	.19	.23*	.15	.14	.06	.10	.12	.29**	.07	.07	.21*	.21*	(.92)

Note. $N = 92$. * $p < .05$, ** $p < .01$. Gender was coded as 1 = male and 2 = female. LGD = leaderless group discussion. Presentation 1 = sales presentation, Presentation 2 = leisure activity presentation, LGD 1 = staffing task, LGD 2 = graduate marketing task, LGD 3 = work-life-balance task. Cronbach's α is reported in parentheses. Cooperation was rated in one exercise only, therefore, no Cronbach's α is reported in this case.

Table 4 shows the correlation matrix with the correlations between all dimension ratings from all exercises. Table 5 reports the same dimension-different exercise correlations and different dimensions-same exercise correlations and thus allows conclusions concerning the construct-related validity of the AC. Ratings of the same dimension across exercises correlated substantially with each other (mean correlation r =.36, p < .01). However, correlations between dimension ratings within exercises were even larger (r = .55, p < .01), indicating that discrimination among dimensions (i.e., discriminant validity) was poor. These results are consistent with previous findings (e.g., Melchers et al., 2007; Woehr & Arthur, 2003) and indicate that the AC used in our study was comparable with other ACs with regard to construct-related validity.

Table 4

Means, Standard Deviations, and Correlations Between Dimension Ratings

Exercise/Dimensions	*M*	*SD*	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.
Presentation 1																		
1. Analytical skills	3.53	0.96																
2. Persuasiveness	3.52	1.07	.72**															
3. Organizing and planning	3.50	0.98	.62**	.63**														
4. Presentation skills	3.38	0.99	.56**	.66**	.68**													
Presentation 2																		
5. Persuasiveness	3.58	0.84	.20	.29**	.30**	.18												
6. Organizing and planning	3.38	1.00	.32**	.45**	.44**	.33**	.55**											
7. Presentation skills	3.45	0.93	.27**	.37**	.37**	.55**	.34**	.55**										
LGD 1																		
8. Analytical skills	2.80	0.88	.21*	.26*	.25*	.28**	.10	.11	.14									
9. Persuasiveness	2.98	0.84	.21*	.38**	.29**	.31**	.18	.37**	.34**	.55**								
10. Organizing and planning	2.89	1.05	.15	.30**	.35**	.32**	.32**	.35**	.35**	.55**	.49**							
11. Cooperation	3.14	0.89	.08	.19	.25*	.13	.21*	.15	.21*	.47**	.47**	.55**						
LGD 2																		
12. Persuasiveness	3.48	0.88	.40**	.34**	.35**	.41**	.19	.31**	.29**	.28**	.43**	.33**	.05					
13. Organizing and planning	3.17	1.06	.31**	.40**	.42**	.42**	.21*	.40**	.34**	.30**	.51**	.41**	.27**	.47**				
14. Assertiveness	2.98	1.02	.28**	.38**	.31**	.30**	.26*	.38**	.26*	.30**	.50**	.34**	.17	.64**	.61**			
LGD 3																		
15. Persuasiveness	3.52	0.75	.19	.18	.27**	.14	.23*	.16	.16	.37**	.44**	.28**	.29**	.42**	.34**	.43**		
16. Organizing and planning	3.03	1.02	.16	.26*	.35**	.30**	.20	.38**	.33**	.40**	.46**	.48**	.43**	.28**	.36**	.32**	.33**	
17. Assertiveness	3.01	1.01	.20	.20	.21*	.18	.31**	.27**	.21*	.35**	.39**	.35**	.33**	.44**	.39**	.49**	.54**	.48**

Note. $N = 92$. * $p < .05$, ** $p < .01$. LGD = leaderless group discussion. Presentation 1 = sales presentation, Presentation 2 = leisure activity presentation, LGD 1 = staffing task, LGD 2 = graduate marketing task, LGD 3 = work-life-balance task.

Table 5

Construct-Related Validity

Dimensions / Exercises	*r*
Same Dimension-Different Exercise Correlations	
Analytical skills	.21*
Persuasiveness	.31**
Organizing and planning	.39**
Assertiveness	–
Cooperation	.50**
Presentation skills	.55**
Mean	**.36****
Different Dimension-Same Exercise Correlations	
Presentation 1	.65**
Presentation 2	.47**
LGD 1	.52**
LGD 2	.58**
LGD 3	.45**
Mean	**.55****

Note. $N = 92$. * $p < .05$, ** $p < .01$. LGD = leaderless group discussion. Presentation 1 = sales presentation, Presentation 2 = leisure activity presentation, LGD 1 = staffing task, LGD 2 = graduate marketing task, LGD 3 = work-life-balance task. Assertiveness was evaluated in one exercise only, therefore, no same dimension-different exercise correlation is reported for this dimension.

Effects of Exercise Similarity on AC Construct-Related and Criterion-Related Validity

We expected convergent validity to be better when exercises are similar compared to when they are dissimilar. To test this assumption, we determined the mean convergent and discriminant validity for each pair of similar and dissimilar exercises, respectively, on the basis of the correlation matrix presented in Table 4. Furthermore, we analyzed the criterion-related validity of similar and dissimilar pairs of exercises, respectively, in two ways: (1) For mean dimension ratings, meaning that we calculated mean ratings on a specific dimension across pairs of exercises and then determined the criterion-related validity for each mean dimension rating obtained, and (2) for the overall rating across pairs of exercises, meaning that we calculated a mean rating across dimensions for each exercise, averaged the respective means across pairs of exercises to obtain overall ratings across pairs of exercises, and then determined the criterion-related validity for each overall rating across pairs of exercises obtained. We averaged the obtained convergent and criterion-related validities, respectively, once across all similar and once across all dissimilar pairs of exercises. All correlations were *r*-to-*Z* transformed prior to averaging.

We conducted all analyses concerning construct-related and criterion-related validity of similar and dissimilar exercises twice: First, we used all the dimensions that were evaluated in the AC exercises. This approach represents conventional AC practice. And second, we used the subset of dimensions that were common to all exercises, namely, organizing and planning, and persuasiveness. In doing so, we answered the demand for holding constructs constant when comparing methods (cf. Arthur & Villado, 2008) and prevented validity coefficients from being influenced by differences in the predictive power of specific dimensions that were rated in some exercises only (cf. Arthur et al., 2003; Meriac et al., 2008).

Mean construct-related and criterion-related validity coefficients of similar and dissimilar exercises are reported in Table 6. When all dimensions were used for the analysis,

convergent validity was significantly better for similar than for dissimilar pairs of exercises, mean r = .44 vs. .31, $t(21) = 3.60$, $p < .01$. However, we found no differences for similar and dissimilar pairs of exercises regarding criterion-related validity on the dimension-level or for mean overall exercise ratings, both ts < 1. Analyses for the two dimensions that were used in all exercises revealed a similar picture: While convergent validity was better for similar than for dissimilar pairs of exercises, mean r = .41 vs. .32, $t(17.99) = 2.88$, $p < .05$, criterion-related validities of similar and dissimilar pairs of exercises did not differ, both ts < 1. Finally, with regard to discriminant validity we found no differences between similar and dissimilar exercises with either set of dimensions, both ts < 1.

Table 6

Mean Construct-Related and Criterion-Related Validities for Similar and Dissimilar Pairs of Exercises

	Construct-related validity				Criterion-related validity			
	Convergent	*k*	Discriminant	*k*	For mean dimension ratings across pairs of exercises	*k*	For overall ratings across pairs of exercises	*k*
All dimensions								
Similar exercises	$.44_a$	10	.54	33	.12	17	.16	4
Dissimilar exercises	$.31_b$	13	.55	51	.14	28	.20	6
Subset of dimensions common to all exercises								
Similar exercises	$.41_a$	8	.47	8	.16	8	.19	4
Dissimilar exercises	$.32_b$	12	.50	12	.18	12	.21	6

Note. N = 92. k = number of correlations included in the calculation of the mean validity coefficient. Different subscripts in a column indicate significant differences between validity coefficients, $p < .05$.

Taken together, with both sets of dimensions, average convergent validity was significantly better for similar than for dissimilar exercises, which is in line with our expectation. However, mean criterion-related validities of pairs of similar exercises did not differ from mean criterion-related validities of pairs of dissimilar exercises.

Discussion

The present study investigated the effects of exercise similarity on both the construct-related and criterion-related validity of an AC. In line with our assumption, the convergent validity of dimension ratings was better when exercises were similar compared to when exercises were dissimilar, indicating that convergent validity of dimension ratings depends on exercise similarity. This is consistent with previous findings (Highhouse & Harris, 1993; Sackett & Harris, 1988; Schneider & Schmitt, 1992) as well as with TAT (Tett & Burnett, 2003; Tett & Guterman, 2000) and implies that exercise similarity seems to allow candidates to show consistent behavior across exercises. Conversely, candidates seem to perform less consistently across dissimilar exercises (cf. Highhouse & Harris, 1993), which is in line with the assumption that different exercises elicit different behaviors (cf. Howard, 2008; Neidig & Neidig, 1984; see also Sackett & Harris, 1988).

Furthermore, we intended to investigate whether exercise similarity also influences the criterion-related validity of dimension ratings. Our results showed that the criterion-related validity of ratings from similar exercises was not significantly different from the criterion-related validity of ratings from dissimilar exercise, suggesting that exercise similarity had no effect on AC criterion-related validity.

Thus, our results imply that improvements in construct-related validity are not necessarily paralleled by improvements in criterion-related validity. A possible explanation for this finding is that our manipulation potentially influenced two variables: First,

candidates' behavioral consistency and second, – as a consequence – the reliability of the mean ratings across exercises (i.e., mean dimension ratings and overall ratings across pairs of exercises, both obtained by averaging ratings across exercises). When exercises were similar, candidates' behavior could be evaluated in more similar situations than when exercises were dissimilar. Thus, the increased similarity of the exercises potentially led to more similar behavioral reactions, that is, to more behavioral consistency and thereby to greater convergence of the dimension ratings across exercises when exercises were similar compared to when they were dissimilar. At the same time, focusing on multiple similar exercises resembles the situation of increasing the reliability of a measure by adding additional parallel items, so that the reliability of mean ratings across exercises was possibly better when exercises were similar compared to when exercises were dissimilar (cf. Brannick, 2008). The improved reliability of the mean ratings across exercises would be beneficial for criterion-related validity. However, the increased behavioral consistency associated with greater exercise similarity might have restricted the range of observed behaviors possibly relevant for job performance. This, in turn, would be disadvantageous for criterion-related validity (cf. Gaugler et al., 1987). Thus, exercise similarity might have had two different effects on criterion-related validity that simultaneously offset each other. As a result, criterion-related validity of similar and dissimilar exercises did not differ but seemed to remain unaffected by exercise similarity.

Our results are inconsistent with findings from previous studies that found improvements in construct-related validity to be paralleled by improvements in criterion-related validity (Melchers et al., 2010; Schleicher et al., 2002). However, those previous studies selectively manipulated factors that affected the reliability and accuracy of single dimension ratings within each exercise. For example, frame-of-reference training led to more reliable and more accurate dimension ratings than control training (Schleicher et al., 2002). In

contrast, in the present study, exercise similarity was expected to influence candidates' behavioral consistency across exercises. In sum, this suggests that only factors that selectively influence the reliability and accuracy of dimension ratings at the level of individual exercises should have similar effects on construct-related and criterion-related validity of ACs that are in line with the unitarian framework of validity (e.g., Binning & Barrett, 1989; Landy, 1986; Messick, 1995).

Practical Implications

Our findings revealed that convergence between ratings on specific dimensions across exercises will more likely be established when exercises are similar compared to when exercises are dissimilar. This implies that the overall convergent validity coefficient of an AC consisting of different exercises potentially provides a too negative picture and that it might be premature to denounce such ACs as not being construct valid. Furthermore, in light of our study, findings concerning the construct-related validity of ACs cannot be generalized to all ACs, because each AC is individually designed (i.e., each AC has a larger or smaller number of similar and dissimilar exercises). Based on these conclusions, one way to obtain a more appropriate estimate for the convergent validity of AC ratings would be to consider only similar exercises. Thus, convergent validity could be determined separately for each type of exercise. Furthermore, when conducting ACs for purposes for which construct-related validity is particularly important (e.g., for developmental purposes), one option would be to use sets of exercises that pose similar demands on candidates and thus allow to assess the consistency of candidates' behavior (i.e., using multiple exercises of the same type; see also Brannick, 2008). In contrast, our results suggest that for selection purposes for which the prediction of job performance is of particular interest it makes no difference whether the AC consists of similar or dissimilar exercises. However, candidates will probably perceive the AC to be more fair

when they have the opportunity to perform different tasks in which an appropriate evaluation of a broad range of job-related behaviors is possible (cf. Bertolino & Steiner, 2007; Gilliland, 1993). Therefore, when conducting ACs for selection purposes, it might be more important to use exercises that represent diverse job-related situations than it is to focus on similar exercises.

Findings from the present study also suggest that improvements in construct-related validity are not always paralleled by improvements in criterion-related validity. Therefore, AC users should not be misguided by the belief that interventions to improve one aspect of validity are also necessarily beneficial for other aspects of validity.

Limitations and Suggestions for Future Research

The present study has several limitations that need to be addressed. First, we used a simulated graduate AC. However, almost all participants indicated that they acted as they would in a real selection situation. Second, the participants' jobs for which criterion data were obtained were heterogeneous. Although the AC was designed in such a way that it covered requirements that are essential in many graduate jobs, the heterogeneity of the participants' jobs might have impaired the criterion-related validity of the AC.

Furthermore, we operationalized exercise similarity through exercise type (cf. Schneider & Schmitt, 1992) and focused on leaderless group discussions and presentation exercises. Future research should investigate whether our findings generalize to other types of exercises, for example, to role plays and case studies.

Even though we did not find parallel effects of exercise similarity on construct-related and criterion-related validity of an AC, exercise similarity also did not have opposite effects on construct-related and criterion-related validity of an AC as might be concluded on the basis of previous studies that focused on only one aspect of validity. This finding points out the

importance of examining the effects that AC interventions have on both construct-related and criterion-related validity simultaneously (cf. Lievens, 2009; Lievens et al., 2009; Melchers & König, 2008; Neidig & Neidig, 1984; Woehr & Arthur, 2003). Therefore, our study should encourage others to address both construct-related and criterion-related validity of AC simultaneously to obtain further insight into the connection between different aspects of validity.

References

Arthur, W., Jr., & Day, E. A. (2010). Assessment centers. In S. Zedeck (Ed.), *APA handbook of industrial and organizational psychology: Selecting and developing members for the organization* (Vol. 2, pp. 205-235). Washington, DC: APA.

Arthur, W., Jr., Day, E. A., McNelly, T. L., & Edens, P. S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology, 56*, 125-154. doi:10.1111/j.1744-6570.2003.tb00146.x

Arthur, W., Jr., & Villado, A. J. (2008). The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. *Journal of Applied Psychology, 93*, 435-442. doi:10.1037/0021-9010.93.2.435

Becker, N., Höft, S., Holzenkamp, M., & Spinath, F. M. (2011). The predictive validity of assessment centers in German-speaking regions: A meta-analysis. *Journal of Personnel Psychology, 10*, 61-69. doi:10.1027/1866-5888/a000031

Bertolino, M., & Steiner, D. D. (2007). Fairness reactions to selection methods: An Italian study. *International Journal of Selection and Assessment, 15*, pp. doi:10.1111/j.1468-2389.2007.00381.x

Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology, 74*, 478-494. doi:10.1037/0021-9010.74.3.478

Bowler, M. C., & Woehr, D. J. (2006). A meta-analytic evaluation of the impact of dimension and exercise factors on assessment center ratings. *Journal of Applied Psychology, 91*, 1114-1124. doi:10.1037/0021-9010.91.5.1114

Brannick, M. T. (2008). Back to basics of test construction and scoring. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 131-133. doi:10.1111/j.1754-9434.2007.00025.x

Chan, D. (1996). Criterion and construct validation of an assessment centre. *Journal of Occupational and Organizational Psychology, 69*, 167-181. doi:10.1111/j.2044-8325.1996.tb00608.x

Dilchert, S., & Ones, D. S. (2009). Assessment center dimensions: Individual differences correlates and meta-analytic incremental validity. *International Journal of Selection and Assessment, 17*, 254-270. doi:10.1111/j.1468-2389.2009.00468.x

Eurich, T. L., Krause, D. E., Cigularov, K., & Thornton, G. C., III. (2009). Assessment centers: Current practices in the United States. *Journal of Business and Psychology, 24*, 387-407. doi:10.1007/s10869-009-9123-3

Fleenor, J. W. (1996). Constructs and developmental assessment centers: Further troubling empirical findings. *Journal of Business and Psychology, 10*, 319-335. doi:10.1007/BF02249606

Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., III, & Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology, 72*, 493-511. doi:10.1037/0021-9010.72.3.493

Gilliland, S. W. (1993). The perceived fairness of selection systems: An organizational justice perspective. *The Academy of Management Review, 18*, 694-734. doi:10.2307/258595

Goffin, R. D., Jelley, R. B., Powell, D. M., & Johnston, N. G. (2009). Taking advantage of social comparisons in performance appraisal: The relative percentile method. *Human Resource Management, 48*, 251-268. doi:10.1002/hrm.20278

Haaland, S., & Christiansen, N. D. (2002). Implications of trait-activation theory for evaluating the construct validity of assessment center ratings. *Personnel Psychology, 55*, 137-163. doi:10.1111/j.1744-6570.2002.tb00106.x

Hardison, C. M., & Sackett, P. R. (2004). *Assessment center criterion related validity: A meta-analytic update*. Unpublished manuscript.

Henderson, F., Anderson, N., & Rick, S. (1995). Future competency profiling: Validating and redesigning the ICL graduate assessment centre. *Personnel Review, 24*, 19-31. doi:10.1108/00483489510089614

Hermelin, E., Lievens, F., & Robertson, I. T. (2007). The validity of assessment centres for the prediction of supervisory performance ratings: A meta-analysis. *International Journal of Selection and Assessment, 15*, 405-411. doi:10.1111/j.1468-2389.2007.00399.x

Highhouse, S., & Harris, M. M. (1993). The measurement of assessment center situations: Bem's template matching technique for examining exercise similarity. *Journal of Applied Social Psychology, 23*, 140-155. doi:10.1111/j.1559-1816.1993.tb01057.x

Hoffman, B., Melchers, K. G., Blair, C. A., Kleinmann, M., & Ladd, R. T. (2011). Exercises AND dimensions are the currency of assessment centers. *Personnel Psychology, 64*, 351-395. doi:10.1111/j.1744-6570.2011.01213.x

Howard, A. (2008). Making assessment centers work the way they are supposed to. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 98-104. doi:10.1111/j.1754-9434.2007.00018.x

International Task Force on Assessment Center Guidelines. (2009). Guidelines and ethical considerations for assessment center operations. *International Journal of Selection and Assessment, 17*, 243-253. doi:10.1111/j.1468-2389.2009.00467.x

Jansen, P. G., & Stoop, B. A. (2001). The dynamics of assessment center validity: Results of a 7-year study. *Journal of Applied Psychology, 86*, 741-753. doi:10.1037/0021-9010.86.4.741

Krause, D. E., & Gebert, D. (2003). A comparison of assessment center practices in organizations in german-speaking regions and the United States. *International Journal of Selection and Assessment, 11*, 297-312. doi:10.1111/j.0965-075X.2003.00253.x

Krause, D. E., Kersting, M., Heggestad, E. D., & Thornton, G. C., III. (2006). Incremental validity of assessment center ratings over cognitive ability tests: A study at the executive management level. *International Journal of Selection and Assessment, 14*, 360-371. doi:10.1111/j.1468-2389.2006.00357.x

Krause, D. E., & Thornton, G. C., III. (2009). A cross-cultural look at assessment center practices: Survey results from Western Europe and North America. *Applied Psychology: An International Review, 58*, 557-585. doi:10.1111/j.1464-0597.2008.00371.x

Lance, C. E., Lambert, T. A., Gewin, A. G., Lievens, F., & Conway, J. M. (2004). Revised estimates of dimension and exercise variance components in assessment center postexercise dimension ratings. *Journal of Applied Psychology, 89*, 377-385. doi:10.1037/0021-9010.89.2.377

Landy, F. J. (1986). Stamp collecting versus science: Validation as hypothesis testing. *American Psychologist, 41*, 1181-1192. doi:10.1037/0003-066X.41.11.1183

Lievens, F. (2009). Assessment centres: A tale about dimensions, exercises, and dancing bears. *European Journal of Work and Organizational Psychology, 18*, 102-121. doi:10.1080/13594320802058997

Lievens, F., Chasteen, C. S., Day, E. A., & Christiansen, N. D. (2006). Large-scale investigation of the role of trait activation theory for understanding assessment center convergent and discriminant validity. *Journal of Applied Psychology, 91*, 247-258. doi:10.1037/0021-9010.91.2.247

Lievens, F., Dilchert, S., & Ones, D. S. (2009). The importance of exercise and dimension factors in assessment centers: Simultaneous examinations of construct-related and criterion-related validity. *Human Performance, 22*, 375-390. doi:10.1080/08959280903248310

Melchers, K. G., & Annen, H. (2010). Officer selection for the Swiss armed forces: An evaluation of validity and fairness issues. *Swiss Journal of Psychology, 69*, 105-115. doi:10.1024/1421-0185/a000012

Melchers, K. G., Henggeler, C., & Kleinmann, M. (2007). Do within-dimension ratings in assessment centers really lead to improved construct validity? A meta-analytic reassessment. *Zeitschrift für Personalpsychologie, 6*, 141-149. doi:10.1026/1617-6391.6.4.141

Melchers, K. G., Kleinmann, M., & Prinz, M. A. (2010). Do assessors have too much on their plates? Rating quality and the number of simultaneously observed candidates in assessment center group discussions. *International Journal of Selection and Assessment, 18*, 329-341. doi:10.1111/j.1468-2389.2010.00516.x

Melchers, K. G., & König, C. J. (2008). It is not yet time to dismiss dimensions in assessment centers. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 125-127. doi:10.1111/j.1754-9434.2007.00023.x

Meriac, J. P., Hoffman, B. J., Woehr, D. J., & Fleisher, M. S. (2008). Further evidence for the validity of assessment center dimensions: A meta-analysis of the incremental criterion-related validity of dimension ratings. *Journal of Applied Psychology, 93*, 1042-1052. doi:10.1037/0021-9010.93.5.1042

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749. doi:10.1037/0003-066X.50.9.741

Neidig, R. D., & Neidig, P. J. (1984). Multiple assessment center exercises and job relatedness. *Journal of Applied Psychology, 69*, 182-186. doi:10.1037/0021-9010.69.1.182

Sackett, P. R. (1987). Assessment centers and content validity: Some neglected issues. *Personnel Psychology, 40*, 13-25. doi:10.1111/j.1744-6570.1987.tb02374.x

Sackett, P. R., & Dreher, G. F. (1982). Constructs and assessment center dimensions: Some troubling empirical findings. *Journal of Applied Psychology, 67*, 401-410. doi:10.1037/0021-9010.67.4.401

Sackett, P. R., & Harris, M. M. (1988). A further examination of the constructs underlying assessment center ratings. *Journal of Business and Psychology, 3*, 214-229. doi:10.1007/BF01014490

Schleicher, D. J., Day, D. V., Mayes, B. T., & Riggio, R. E. (2002). A new frame for frame-of-reference training: Enhancing the construct validity of assessment centers. *Journal of Applied Psychology, 87*, 735-746. doi:10.1037/0021-9010.87.4.735

Schneider, J. R., & Schmitt, N. (1992). An exercise design approach to understanding assessment center dimension and exercise constructs. *Journal of Applied Psychology, 77*, 32-41. doi:10.1037/0021-9010.77.1.32

Staufenbiel, T., & Hartz, C. (2000). Organizational citizenship behavior: Development and validation of a measurement instrument. *Diagnostica, 46*, 73-83. doi:10.1026//0012-1924.46.2.73

Tett, R. P., & Burnett, D. D. (2003). A personality trait-based interactionist model of job performance. *Journal of Applied Psychology, 88*, 500-517. doi:10.1037/0021-9010.88.3.500

Tett, R. P., & Guterman, H. A. (2000). Situation trait relevance, trait expression, and cross-situational consistency: Testing a principle of trait activation. *Journal of Research in Personality, 34*, 397-423. doi:10.1006/jrpe.2000.2292

Thornton, G. C., III, & Rupp, D. E. (2006). *Assessment centers in human resource management: Strategies for prediction, diagnosis, and development*. Mahwah, NJ: Lawrence Erlbaum.

Woehr, D. J., & Arthur, W., Jr. (2003). The construct-related validity of assessment center ratings: A review and meta-analysis of the role of methodological factors. *Journal of Management, 29*, 231-258. doi:10.1177/014920630302900206

Chapter 3

The Relation Between Assessment Center Overall Dimension Ratings and External Ratings of the Same Dimensions

Andreja Wirz[1], Klaus G. Melchers[1], Martin Kleinmann[1], Filip Lievens[2], Hubert Annen[3], and Urs Bettler[1]

[1]Universität Zürich, Switzerland; [2]Ghent University, Belgium; [3]ETH Zürich, Switzerland

Acknowledgements

We thank Sabrina Engeli, Pascale Gschwend, and Stefan Schultheiss for their help with data collection.

Manuscript submitted for publication to *Personnel Psychology*

Abstract

We examined whether relating AC overall dimension ratings to ratings of identical dimensions that stem from sources external to the AC will provide evidence for construct-related validity of ACs. For this purpose, we analyzed data from three samples, two of which were from field settings (*N*s = 428 and 121) and one from a laboratory setting (*N* = 92). Thereby, supervisors, customers, and candidates themselves, respectively, represented external sources. Results showed that different dimension-same source correlations within the ACs were larger than same dimension-different source correlations. Moreover, in all three samples confirmatory factor analyses revealed source factors but no dimension factors in the latent factor structure of overall dimension ratings from the AC and from external sources. These results indicate that AC overall dimension ratings and external dimensions ratings cannot be attributed to the purported dimensions, meaning that relating AC overall dimension ratings to external ratings of identical dimensions is not successful in evidencing AC construct-related validity. However, our findings suggest that ACs and other sources capture different aspects of behavior and that they provide different perspectives on people's performance. Implications for practice and suggestions for research that can be derived from these findings are discussed.

It has been repeatedly shown that assessment centers (ACs) predict future performance and that they have incremental validity beyond cognitive ability tests or personality inventories (cf. Dilchert & Ones, 2009; Krause, Kersting, Heggestad, & Thornton, 2006; Melchers & Annen, 2010; Meriac, Hoffman, Woehr, & Fleisher, 2008). In contrast, findings concerning the internal construct-related validity of ACs are less promising, because they cast some doubts on the degree to which ACs measure the dimensions they are intended to measure (e.g., Bowler & Woehr, 2006; Woehr & Arthur, 2003).

The internal construct-related validity of ACs is usually evaluated on the basis of dimension ratings obtained after the completion of each exercise (within-exercise dimension ratings), where it is commonly found that correlations between ratings of the same dimension across exercises are low and correlations between ratings of different dimensions within exercises are high (e.g., Bowler & Woehr, 2006; Woehr & Arthur, 2003). However, several authors have raised concerns about the appropriateness of within-exercise dimension ratings as the basis for the construct-related validation of ACs (e.g., Neidig & Neidig, 1984; Reilly, Henry, & Smither, 1990; Rupp, Thornton, & Gibbons, 2008). They argued that using within-exercise dimension ratings leads to a misinterpretation of information from ACs because different exercises are not parallel measurement methods that all capture a dimension in exactly the same manner (Howard, 2008; Lievens & Conway, 2001; Neidig & Neidig, 1984). For that reason, they proposed examining the external construct-related validity instead of the internal construct-related validity of ACs by focusing on overall dimension ratings that reflect the overall performance on a dimension. According to this view, overall dimension ratings could be related to other evaluations of the same dimensions that stem from sources external to the AC (e.g., Rupp et al., 2008). Examples of such external sources include other tests and inventories, peer ratings, supervisor ratings, self-ratings, customer ratings, etc.

There are two competing views with regard to the expectation to find evidence for AC construct-related validity when relating AC overall dimension ratings to external ratings of the same dimensions. On the one hand, some previous studies found promising results for the external construct-related validity of ACs. Generally, they demonstrated that AC dimension ratings correlate more with conceptually related constructs gathered from external sources than with conceptually unrelated constructs from these external sources (e.g., Shore, Thornton, & Shore, 1990; Thornton, Tziner, Dahan, Clevenger, & Meir, 1997). On the other hand, research on multisource feedback has found that dimension factors account for less variance in dimension ratings from different sources compared to source factors (e.g., Hoffman, Lance, Bynum, & Gentry, 2010). This suggests that relating AC overall dimension ratings to external ratings of the same dimensions might not be successful in evidencing that ACs measure the purported constructs. However, as explained below, the previous studies that used an external construct-related validation approach do not allow definite conclusions about the relation between AC overall dimension ratings and external evaluations on identical dimensions and thus, which of the two views can be empirically supported. Therefore, it is necessary to investigate how AC overall dimension ratings relate to external ratings of the *same* dimensions.

The aim of the present study was to examine the relation between AC overall dimension ratings and evaluations of identical dimensions that stem from sources external to the AC (cf. Arthur, Day, & Woehr, 2008; Neidig & Neidig, 1984; Reilly et al., 1990; Rupp et al., 2008). We believe that research in this regard is essential for at least three different reasons: First, overall dimension ratings – and not within-exercise dimension ratings – are usually used as the basis for feedback to candidates and to determine employees' developmental needs. Second, overall dimension ratings from ACs provide important information for placement decisions where they are used to determine whether the profile of

candidates' strengths suits the specific demands of one or several different positions. And third, at a conceptual level, the present research will provide an answer to the question of which of the two competing expectations with regard to the presence of substantial dimension variance in AC overall dimension ratings and external ratings of the same dimensions finds more empirical support. That is, we will determine whether AC overall dimension ratings really permit conclusions concerning performance on the specific dimensions and thereby contribute to a more comprehensive insight into the construct-related validity of ACs.

Review of Previous Research

In this section, we will first review research related to the internal construct-related validity of ACs. Then, on the basis of possible explanations for the findings in this regard, we will address the suggestion to examine the external construct-related validity instead of the internal construct-related validity of ACs by relating AC overall dimension ratings to external ratings of the same dimensions, and we will review previous studies. Finally, we will describe relevant research from the multisource feedback domain and research on typical and maximum performance. These two literatures support a more skeptical position concerning the question of whether relating AC overall dimension ratings to external ratings of the same dimensions might provide evidence that ACs measure the dimensions they are designed to measure.

Construct-Related Validity of ACs

There are different approaches to determining the internal construct-related validity of ACs. One of the most frequently used approaches is to compare correlations between ratings of different dimensions. Following this approach, high correlations between ratings of the same dimension that were assessed in different exercises indicate that ratings have convergent

validity. Conversely, low correlations between ratings on different dimensions obtained in the same exercise indicate that ratings have discriminant validity. As a more formal test for examining the construct-related validity of ACs, confirmatory factor analysis (CFA) can be used. ACs are considered to have construct-related validity if the factor structure that underlies dimension ratings incorporates dimension factors that explain an essential proportion of variance in ratings. Both approaches for the internal construct-related validation of ACs usually use dimension ratings obtained after the completion of each exercise, that is, within-exercise dimension ratings, as the basis for analyses.

Many attempts to show evidence for the construct-related validity of ACs have been made. However, research has repeatedly reported problems in finding internal construct-related validity of ACs, suggesting that it is unclear whether ACs measure the intended dimensions. Several studies revealed that different dimension-same exercise correlations are usually high, whereas same dimension-different exercise correlations are low (cf. Melchers, Henggeler, & Kleinmann, 2007; and Woehr & Arthur, 2003, for meta-analytic results). Similarly, CFAs usually revealed that exercise factors represent a more important source of variance of within-exercise dimension ratings than dimension factors – if dimension factors could be found at all (e.g., Bowler & Woehr, 2006; Hoffman, Melchers, Blair, Kleinmann, & Ladd, 2011; Lance, Lambert, Gewin, Lievens, & Conway, 2004; Lievens, Dilchert, & Ones, 2009).

Several explanations for the findings concerning the internal construct-related validity of ACs have been offered. For example, it has been posited that the conventional rating system in ACs introduces common rater variance into within-exercise dimension ratings (Howard, 1997; Kolk, Born, & van der Flier, 2002; Melchers et al., 2007) and that within-exercise dimension ratings are one-item measures that might lack reliability (e.g., Arthur et al., 2008; Howard, 2008). For that reason, it has been suggested to integrate dimension ratings

across exercises into overall dimension ratings that are assumed to reflect candidates' general performance on the dimensions in the AC (cf. Arthur et al., 2008; Neidig & Neidig, 1984; Reilly et al., 1990; Rupp et al., 2008). These overall dimension ratings are expected to be more reliable than within-exercise dimension ratings, and it should be more likely to establish construct-related validity for overall dimension ratings than for within-exercise dimension ratings (cf. Arthur et al., 2008). On this basis, an external construct-related validation approach that uses overall dimension ratings as the focal variables from the AC and dimension evaluations that stem from other sources as comparative data has been proposed (cf. Arthur et al., 2008; Neidig & Neidig, 1984; Reilly et al., 1990; Rupp et al., 2008).

External Construct-Related Validity of ACs

Why an external construct-related validation approach for ACs might be promising. It has repeatedly been emphasized that the theory underlying the conventional construct-related validation approaches does not correspond with the nature of ACs (Howard, 2008; Rupp et al., 2008). Implicit to the approach of analyzing within-exercise dimension ratings for the construct-related validation of ACs is that dimensions should represent stable attributes and exercises should represent different measurement methods that are equally capable of measuring a specific dimension. However, the original idea behind ACs was to use different exercises that allow the assessment of dimensions from different perspectives (Howard, 2008). Consistent with this, different exercises might capture only selected facets of a specific dimension and a specific dimension might be more or less relevant in different exercises (Howard, 2008). Therefore, convergence between ratings on a specific dimension from different exercises is not necessarily to be expected. Similarly, Neidig and Neidig (1984; see also Lievens & Conway, 2001; Lievens, Dilchert et al., 2009) argued that different exercises elicit different behaviors, which might result in cross-situational inconsistency of

candidates' behavior (cf. Lance, Foster, Gentry, & Thoresen, 2004; Lance et al., 2000) and thus explain the low convergence between dimension ratings from different exercises.

This perspective is based on the assumption that behavior is determined by the interaction between person and situation variables. According to interactionist theories (e.g., Mischel & Shoda, 1995; Tett & Burnett, 2003; Tett & Guterman, 2000), people behave differently across situations and thus also in different AC exercises (Lievens, Tett, & Schleicher, 2009; Melchers, Wirz, & Kleinmann, in press). Hence, variation in candidates' behavior across exercises is not necessarily indicative of measurement error (cf. Neidig & Neidig, 1984) or – like Howard (2008) argued more precisely – it might reflect "valuable information, not a broken method" (p. 103). As a consequence, dimension ratings from different exercises should not be treated as parallel measures because this seems to lead to a misinterpretation of AC ratings. Accordingly, researchers (Neidig & Neidig, 1984; see also, for example, Arthur et al., 2008; Reilly et al., 1990; Rupp et al., 2008) proposed not to examine the internal structure of ACs on the basis of within-exercise dimension ratings for construct-related validation but rather to investigate the external construct-related validity of ACs by focusing on overall dimension ratings. It has repeatedly been shown that these overall dimension ratings predict job performance (e.g., Arthur, Day, McNelly, & Edens, 2003; Dilchert & Ones, 2009; Meriac et al., 2008), indicating that they reflect meaningful variance. Therefore, overall dimension ratings should have construct-related validity that might be evidenced when relating them to other evaluations on the same dimensions that stem from other assessment methods, for example, multisource feedback ratings (Rupp et al., 2008).

A few initial studies analyzed dimension ratings from ACs in relation to externally assessed variables and provided evidence for the external construct-related validity of ACs. For example, Shore et al. (1990) correlated overall dimension ratings from an AC with external measures of cognitive ability and personality, respectively. They found that

dimensions classified into a broader performance-style dimension correlated more strongly with measures of cognitive ability than dimensions classified into a broader interpersonal-style dimension. Furthermore, correlations between dimensions and conceptually similar personality measures tended to be higher than correlations between dimensions and conceptually dissimilar personality measures. Similarly, Thornton et al. (1997) found that AC dimension ratings correlated more strongly with conceptually related test measures than with conceptually unrelated test measures. Similar findings were reported by Dilchert and Ones (2009). They found that the correlation between the primary AC dimensions (classified according to suggestions from Arthur et al., 2003) on the one hand, and cognitive ability and specific personality traits, respectively, on the other hand, were higher when the AC dimensions were conceptually related to cognitive ability and to different personality traits than when they were not. In contrast to this, a few other studies called the external construct-related validity of ACs into question. For example, Chan (1996) and Fleenor (1996) did not find higher correlations between dimensions and conceptually related constructs that were assessed external to the AC than between dimensions and conceptually unrelated constructs that were externally assessed.

The aforementioned studies provided important contributions to the understanding of the nomological network of ACs. However, the externally assessed constructs in these studies were typically not directly comparable to the AC dimensions. Instead, AC ratings were related to cognitive ability and personality measures, for example (e.g., Dilchert & Ones, 2009; Shore et al., 1990). Thus, the external comparison scores did not represent external evaluations of the *same* dimensions that were used in the ACs. That is, constructs were not held constant when comparing methods (cf. Arthur & Villado, 2008), meaning that constructs and methods were confounded. Therefore, we still do not know whether external construct-

related validity of ACs can be established when relating AC overall dimension ratings to external ratings of the *same* dimensions.

So far, we are aware of only one study that directly examined the relationship between dimension ratings from an AC and evaluations on the same dimensions used in the AC that stem from other assessment methods. Shore et al. (1992) reported correlations between overall dimension ratings from an AC and peer- and self-evaluations of candidates' performance in the AC. They found that dimension evaluations from the three different sources converged. Furthermore, correlations between ratings of different dimensions provided by the same source were lower than correlations between ratings of the same dimension provided by different sources. Results of this study supported the external construct-related validity of ACs overall dimension ratings. However, the peer- and self-evaluations of candidates' performance were incorporated into AC overall dimension ratings, meaning that the AC overall dimension ratings were in part based on the external comparison scores. Hence, the results might also be influenced by a lack of independence between assessment methods. Furthermore, Shore et al. did not evaluate the latent factor structure of dimension ratings.

Taken together, the scarce number of studies on the external construct-related validity of ACs seems to offer some supporting results. However, these studies also have important limitations. That is, either they compared different constructs assessed with different methods, or the methods were not independent of each other (e.g., Shore et al., 1992), which might have influenced the results obtained. Therefore, these previous studies do not allow definite conclusions to be made concerning the relation between AC overall dimension ratings and ratings of the same dimensions provided by sources external to the AC. Thus, the need to investigate the external construct-related validity of ACs as proposed by several authors persists (e.g., Arthur et al., 2008; Neidig & Neidig, 1984; Reilly et al., 1990; Rupp et al.,

2008). In light of the findings in the AC domain reported in this section, it seems reasonable to expect that evidence for the external construct-related validity of an AC can be established when relating overall dimension ratings from an AC to evaluations of the same dimensions that stem from sources external to the AC.

Why an external construct-related validation approach for ACs might not be promising. In the previous section, we explained why it might be possible to find evidence supporting the external construct-related validity of ACs by relating AC overall dimension ratings to external ratings of the same dimensions. However, there are also reasons to expect less promising results for this construct-related validation approach. When examining the external construct-related validity of ACs (e.g., Arthur et al., 2008; Neidig & Neidig, 1984; Reilly et al., 1990; Rupp et al., 2008), dimension evaluations from different assessment methods or different sources are needed, just like in multisource feedback (Rupp et al., 2008). Examples of these different sources are the candidates themselves, peers, customers, subordinates, or supervisors. When evaluating candidates' behavior, each of the latter four will probably refer to situations experienced with the candidate.

As argued above, interactionist theories posit that the interaction between person and situation variables determines behavior (Mischel & Shoda, 1995; Tett & Burnett, 2003; Tett & Guterman, 2000). Therefore, a person probably behaves differently depending on whether he or she interacts with a supervisor, a peer, a subordinate, or a customer. As a consequence, different sources observe different behavior and they have different, relatively unique perspectives on candidates' performance (Borman, 1974; 1997). In addition, different raters capture different slices of behavior and they evaluate behavior differently (Borman, 1974; 1997). For these reasons, ratings on a specific dimension from different sources might not necessarily converge.

Multisource feedback systems take advantage of the unique perspectives of different sources in order to attain a more comprehensive picture of an employee's performance. In line with this, a meta-analysis by Conway and Huffcutt (1997) revealed higher correlations between performance ratings within sources than between performance ratings from different sources, indicating that different sources have different perspectives on performance. Furthermore, several studies found source variance in ratings to be substantial relative to dimension variance (e.g., Conway, 1996; Woehr, Sheehan, & Bennett, 2005). More recently, Hoffman et al. (2010) showed that source factors accounted for a considerably larger proportion of variance in dimension ratings (22% on average) than dimension factors (7% on average). These findings offer support for the unique perspective of different sources. Furthermore, they suggest that the contribution of dimensions to the variance in multisource feedback ratings is limited.

In addition to this, the AC offers a perspective on candidates' behavior that might be different than that of other sources. Specifically, the distinction between maximum and typical performance becomes relevant (Sackett, Zedeck, & Fogli, 1988). The AC reflects a clearly defined assessment situation of limited duration that directly determines the candidates' career. Thus, candidates will increase their effort to perform well in the AC as compared to on the job and they will be more likely to show maximum performance in the AC than in the job context (cf. Marcus, Goffin, Johnston, & Rothstein, 2007; Ployhart, Lim, & Chan, 2001). In contrast, performance on the job reflects typical performance (McCloy, Campbell, & Cudeck, 1994). As a consequence, dimension ratings from an AC might differ from those that refer to the job, which might restrict convergence between AC dimension ratings and dimension ratings from other sources.

Some empirical findings offer support for the distinction between AC performance and performance in the job context in terms of the distinction between typical and maximum

performance. For a military sample, Ployhart et al. (2001), for example, reported that AC ratings and ratings of performance during a military training program correlated less than different ratings from the AC with each other or different measures of performance during the training with each other, respectively.

Taken together, overall dimension ratings from an AC might not converge with dimension ratings from other sources that refer to the job context. In light of previous findings on multisource feedback, source factors will probably explain more variance in AC overall dimension ratings and in dimension ratings from other sources than dimension factors.

The Present Research

We described two competing views on the relation between AC overall dimension ratings and external dimension ratings. On the one hand, there are arguments from the AC literature for expecting that AC overall dimension ratings converge with external ratings of the same dimensions. On the other hand, the multisource feedback literature and research on typical and maximum performance offer arguments against this expectation. However, previous research does not allow definite conclusions in this regard, meaning that it is unclear to what degree AC overall dimension ratings and external ratings of the same dimension converge and whether they can be attributed to dimension factors. Therefore, we aim to extend previous research by evaluating the external construct-related validity of ACs, thereby focusing on the relationship between overall dimension ratings from an AC and external evaluations of the same dimensions, that is, evaluations of the same dimensions that stem from sources external to the AC. This means that we intend to hold constructs constant across methods to ensure that constructs and methods are not confounded (cf. Arthur & Villado, 2008). In addition, we will ensure that the methods used are independent of each other. Of

particular interest is whether this approach to external construct-related validation allows us to find dimension factors that explain a substantial amount of variance in AC dimension ratings.

Method

In the present study, we used three independent samples, two from field settings and one from a laboratory setting. For each sample, dimension ratings from an AC and from two external sources were used. Data from Sample 1 stem from Hagan, Konopaske, and Bernardin (2006), who investigated the criterion-related validity of a multisource performance rating system. In contrast to their study, we focused on the external construct-related validity of dimension ratings from an AC and thus our research represents a reanalysis and extension of Hagan et al.'s analyses. Data from Sample 2 and Sample 3 were gathered for the present study and allow analyses not only with regard to the external construct-related validity, but also with regard to the internal construct-related validity and criterion-related validity of the two ACs. This enabled us to determine whether these ACs were comparable to other ACs in the literature and thus to ensure that findings from the present study are not determined by special characteristics of the ACs considered.

Participants and Procedure

Sample 1. As mentioned above, the data from Sample 1 stem from a study by Hagan, et al. (2006). For the present study, we used selected data from the correlation matrix published in their article (see Table 1). The total sample consisted of 428 associate store managers (71% males, 29% females) from a large retail company who had worked at least one year in the company and who were performing well.

Participants attended a one-day AC for the selection of candidates for promotion to store manager. The AC consisted of an in-basket exercise, two leaderless group discussions, a

case analysis, and an oral presentation that focused on six dimensions critical for performance as (associate) store manager. The six dimensions were oral presentation and communication, written communication (e.g., "clear expression of ideas in writing and in good grammatical form", Hagan et al., 2006, p. 365), interpersonal skills, planning and organizing, decision making, and leadership. Unfortunately, dimension definitions (despite the definition of the dimension written communication) and further information on the exercises were not reported by Hagan et al. and, therefore, cannot be provided here.

Assessors were employees of the retail company who held higher-level positions than the candidates. Prior to the AC, assessors took part in frame-of-reference (FOR) rater training (cf. Woehr & Huffcutt, 1994), where they received specific examples of candidates' performance in the exercises.

In the AC, teams of assessors evaluated the candidates' performance after the completion of all exercises. Each assessor independently rated the dimensions on seven-point behavior expectation scales, with higher numbers indicating better performance. Using the behavior expectation scales, assessors were asked to judge what level of performance on a specific dimension they would expect for a candidate at the store manager level. The behavior expectation scales provided behavioral anchors for different levels of performance for specific dimensions and exercises. More information on the scales can be found in Hagan et al. (2006). Afterwards, assessors met for consensus discussion to determine an overall rating on each dimension for each candidate. These overall dimension ratings, which represented one-item measures, were used for the present analyses.

Two sources external to the AC, namely supervisors and professional customers, evaluated the candidates' performance on the AC dimensions in the same month in which the AC was conducted. Professional customers were mystery shoppers engaged by the retail company who were instructed to act according to scripts. As only 390 AC candidates received

a customer assessment, analyses of the external construct-related validity of the AC are based on $n = 390$. Both supervisors and professional customers used the same seven-point behavior expectation scales that were used in the AC to assess each dimension with one item. However, behavioral anchors were adopted so that they referred to the job situation or to the situation in the scripts that were used for the customer assessment, respectively. More information on the supervisor assessment and customer assessment can be found in Hagan et al. (2006).

Sample 2. Sample 2 consisted of 121 candidates who successfully passed the AC for the selection of prospective career officers in the Swiss Armed Forces between 2003 and 2009 and who were permitted to attend the career officer training in the Swiss Armed Forces. Of these, 116 candidates were male, and only five were female. The candidates' average age at the time of the AC was 27.10 years ($SD = 3.26$).

The AC for the selection of prospective career officers in the Swiss Armed Forces was designed to represent requirements imposed on career officers. In previous studies with other candidates, this AC has been found to have criterion-related validity for both training performance as well as military career success (Gutknecht, Semmer, & Annen, 2005; Melchers & Annen, 2010). Over two days, candidates completed six exercises. In a short oral presentation, each candidate had to introduce him- or herself to the other candidates and assessors and to express his or her opinion on a specific matter. In a leaderless group discussion that represented a problem solving task, candidates were instructed to assert their own interests, while representing the group's interests. In a motivational talk, each candidate had to convince a role player to perform an unpleasant task or to accept a challenging situation. In a debate, candidates first had to agree on a discussion topic and then to convince others of their opinion on this topic. In short case scenarios, each candidate had to describe how he or she would act in three difficult situations that may occur in the everyday life of a

career officer. Finally, candidates had to give a fifteen minute lecture on an aspect of military pedagogy that they could prepare during the spare time between the exercises. In each exercise, candidates were evaluated on three to four dimensions out of six, namely on achievement motivation, analysis, dealing with conflicts, interpersonal skills, oral communication, and influencing others. Dimension definitions and a dimension by exercise matrix can be found in the Appendix (Tables A1 and A2).

The assessor group usually consisted of personnel managers from the Swiss Armed Forces and civilian psychologists or HR experts with experience in personnel selection and assessment. Assessors took part in a one-day rater training session prior to serving as assessors in the AC. In the rater training, assessors received information on ACs and FOR training to practice observing and evaluating (cf. Woehr & Huffcutt, 1994). In addition, directly before the AC assessors participated in a short refresher training.

Candidates were rated by two assessors after each exercise, whereby assessors rotated across exercises. First, assessors independently rated the targeted dimensions on a behaviorally anchored four-point scale (from 1 = *clearly not fulfilled* to 4 = *clearly fulfilled)*. Then, they had to derive a consensus rating for each dimension in the specific exercise. All dimensions were rated in three to five exercises. By averaging the dimension-specific consensus ratings across exercises, we obtained overall dimension ratings for the AC. Coefficient alphas for overall dimension ratings from the AC ranged from .12 to .49 (see Table 2). These results are comparable to previous findings (Atkins & Wood, 2002) and indicate that the internal consistency of the dimension ratings from the AC was low.

To obtain ratings on the AC dimensions from sources external to the AC, a self-evaluation and a supervisory assessment of the AC dimensions were conducted. On average, the time lag between the AC and the external assessment of AC dimensions was 2.55 years (*SD* = 1.38). Supervisors were the candidates' course commanders (i.e., direct military

superiors) who had regular contact with them. Supervisors completed a questionnaire assessing each AC dimension with four items. One of those four items focused on the general performance on the specific dimension based on its definition. The other three items were based on behavioral anchors that were used in the AC and thus focused on specific behaviors related to the dimensions.

According to Goffin, Jelley, Powell, and Johnston (2009), the validity of performance ratings improves when the rating method encourages social comparisons compared to when the rating method is non-comparative. Therefore, instructions as well as the formulation of the items asked supervisors to evaluate the candidates' performance in comparison to other prospective career officers. Ratings were made on a five-point scale, where 1 indicated that the candidate belonged to the poorest performers and 5 indicated that the candidate belonged to the best performers. In the self-evaluation, candidates completed the same questionnaire as the supervisors did. However, the instructions and the items were adapted to be consistent with the candidates' perspective. In both the self-evaluation and the supervisory assessment, we used the statistical means across all items that assessed a specific dimension as external dimension ratings. Coefficient alphas for dimension ratings from the supervisory assessment and the self-evaluation ranged from .95 to .97, and from .64 to .90, respectively (see Table 2).

To examine the criterion-related validity of the AC in addition to the construct-related validity, we used military training performance as criterion. Military training performance referred to the evaluation in the practical military training that alternated with academic training courses. Direct military superiors evaluated the candidates' overall military training performance on a five-point scale (ranging from 1 = *worst* to 5 = *best*).

Sample 3. Sample 3 consisted of 92 recent or prospective university graduates who voluntarily participated in a graduate AC that was administered at a Swiss university for training purposes. Fifty percent of the sample was male and 50% was female. The

participants' average age was 29.10 years (*SD* = 6.20). Almost half of the participants held a Master's degree (47.8%) and 22.8% held a Bachelor's degree. Participants had worked at least 12 hours per week during a six-month period before the AC, mostly in education and research (46.7%), in the banking and insurance industry (10%), or in the service industry (10%).

The one-day graduate AC covered a wide range of requirements essential for a variety of jobs and consisted of five exercises: In a sales presentation, participants had to persuade a potential client of a fictitious company to purchase a manufacturing system. In a second presentation exercise, participants were asked to present a leisure activity of their own choice (e.g., volleyball, hiking, photography, literature, etc.) to a group of other graduates. Another exercise was a leaderless group discussion that represented a staffing task. That is, participants had to identify the best applicant for a vacant position in a fictitious bank. To find the proper solution, participants needed to collaborate and to share previously provided information on the applicants that was distributed among the group. In a second leaderless group discussion, the group had to find a common rank order of ten graduate marketing activities with regard to their efficacy. Prior to this group discussion, participants had to individually determine the perceived efficacy of the graduate marketing activities and they were instructed that the common rank order should correspond as much as possible with their individual view. The last exercise was a leaderless group discussion that was similar to the latter group discussion, but the points of discussion were ten activities for the improvement of the work-life-balance of a company's employees. The evaluations in the exercises referred to three or four dimensions out of six: Analytical skills, persuasiveness, organizing and planning, assertiveness, cooperation, and presentation skills. Dimension definitions and the dimension by exercise matrix can be found in Tables A3 and A4 in the Appendix.

Assessors were Master's level psychology students who took part in a one-day rater training session prior to their first assignment in the AC. The rater training included general information on ACs, an introduction to the dimension definitions and exercises, information on the observation and evaluation process, and FOR training (cf. Woehr & Huffcutt, 1994). Assessors who were not able to participate in the rater training were required to shadow a trained assessor in an AC or to attend an individual training session prior to their first assignment. In the AC, participants were evaluated by rotating teams of two assessors. Directly after having completed an exercise, both assessors independently evaluated the participants' performance on the pre-defined dimensions using a five-point scale (from 1 = *poor* to 5 = *excellent*). After the completion of all exercises, assessors discussed and adjusted dimension ratings that diverged more than one point. The average intraclass correlation of the post-discussion dimension ratings (ICC 1.1), which represents the reliability of a single assessor, was $r = .72$. The averaged post-discussion ratings on specific dimensions across assessors and exercises depicted overall dimension ratings for the AC. Coefficient alphas for overall dimension ratings from the AC ranged from .35 to .76 (see Table 3) and were somewhat higher than found in Atkins and Wood (2002), for example. However, organizing and planning, presentation skills, and persuasiveness were the only three dimensions with an acceptable internal consistency (coefficient alphas of .76, .71, and .69, respectively).

External ratings on the AC dimensions with regard to participants' performance on the job were obtained from two sources, namely from the participants themselves and from their supervisors. For the self-evaluation on the AC dimensions, participants completed seven to eight items per dimension. One of those items directly asked for the overall performance on the specific dimension. The remaining items asked for specific behaviors related to the dimension and were based on the behavioral anchors used in the AC. As in Sample 2, instructions and items in Sample 3 also asked participants to evaluate themselves in

comparison to colleagues in a similar position (cf. Goffin et al., 2009), using a five-point scale with higher numbers indicating better performance. The self-evaluation form for the AC dimensions was administered directly after the completion of the AC but before participants received feedback pertaining to their AC performance. The questionnaire for the supervisory assessment was based on the questionnaire for the self-evaluation. However, the number of items per dimension that focused on specific dimension-relevant behaviors was reduced to four so that supervisors completed five items per dimension. Again, supervisory ratings were made in comparison to the participants' colleagues or in comparison to former employees in a similar position, using a five-point scale (with higher numbers indicating better performance). To obtain dimension ratings from the self-evaluation and the supervisory assessment, respectively, we calculated the statistical mean across all items that assessed a specific dimension. The dimension ratings from the supervisory assessment and the self-evaluation reached coefficient alphas between .70 and .86 and between .74 and .89, respectively.

Furthermore, to examine the criterion-related validity of the AC ratings, the participants' supervisors were asked to evaluate the participants' job performance on five items from the task-based job performance questionnaire by Bott, Svyantek, Goodman, and Bernal (2003) and five items from the German translation (Staufenbiel & Hartz, 2000) of Williams' and Anderson's (1991) in-role behavior scale. Again, ratings were made in comparison to the participants' colleagues or in comparison to former employees in a similar position, using a 7-point scale (with higher numbers indicating better performance). For the analyses, we used the statistical mean across all ten items, which had a coefficient alpha of .92.

Results

Preliminary Analyses

Before examining the external construct-related validity of overall dimension ratings from the ACs, we analyzed the internal construct-related validity of the ACs. For this purpose we calculated the mean correlation between ratings on the same dimension across exercises (i.e., convergent validity), and the mean correlation between ratings on different dimensions within exercises (i.e., discriminant validity). All correlations were *r*-to-*Z* transformed prior to averaging. Furthermore, we determined whether the ACs considered were comparable with other ACs from the literature. For this purpose, we analyzed the criterion-related validity of the ACs by correlating the overall AC ratings, that is, the statistical mean across dimensions and exercises, with job performance criteria.

Sample 1. Dimension ratings from the AC represented single-item measures and, therefore, we were not able to determine the internal construct-related validity of the AC. Furthermore, since no information on job performance was provided in Hagan et al. (2006), we were not able to analyze the criterion-related validity of the AC.

Sample 2. The mean same dimension-different exercise correlation was $r = .12$, and the mean different dimension-same exercise correlation was $r = .33$, indicating that the AC did not have internal construct-related validity. These results are comparable to previous findings on the internal construct-related validity of ACs (e.g., Melchers et al., 2007; Woehr & Arthur, 2003).

Concerning criterion-related validity, we found that the correlation between the overall AC rating and military training performance was $r = .34$, $p < .01$ ($n = 99$). This indicates that the AC was a good predictor of military training performance and comparable to other ACs found in the literature with regard to criterion-related validity (e.g., Becker,

Höft, Holzenkamp, & Spinath, 2011; Gaugler, Rosenthal, Thornton, & Bentson, 1987; Hardison & Sackett, 2004; Hermelin, Lievens, & Robertson, 2007).

Sample 3. The mean same dimension-different exercise correlation was $r = .36$. However, the mean different dimension-same exercise correlation was even larger with $r = .55$, which is problematic with regard to internal construct-related validity.

With regard to criterion-related validity, the correlation between the overall AC rating and job performance was $r = .21$, $p < .05$. This indicates that the present AC had comparable validity for predicting job performance as most ACs (cf. Gaugler et al., 1987; Hardison & Sackett, 2004; Hermelin et al., 2007).

External Construct-Related Validity

We examined the external construct-related validity of the ACs in two ways. First, we compared the mean correlation of ratings of a specific dimension across different sources to the mean correlation of ratings of various dimensions within the AC. Before averaging, all correlations were *r*-to-*Z* transformed. Second, we conducted CFAs to examine the latent factor structure of dimension ratings from different sources.

On the basis of previous research on AC construct-related validity, we tested three sets of models: The first set of models contained conventional models that are comparable to models usually used for construct-related validation of ACs. The model with correlated dimensions (CD-model) hypothesized that only dimension factors determine dimension ratings from the AC and from other sources. The model with correlated sources (CS-model) proposed that candidates' behavior is situationally specific or, in other words, that different sources capture different aspects of candidates' general performance (Borman, 1974; 1977; Conway, 1996; Woehr et al., 2005). The third model in this set comprised both correlated dimensions and correlated sources (CDCS-model).

In the second set of models, we tested models with a general performance factor that suggests that all dimension ratings are based on candidates' overall performance effectiveness (cf. Lance, Foster et al., 2004; Lance et al., 2000). Specifically, we tested a model with only a general performance factor (1G-model). This model proposed that different sources have similar perceptions of candidates' overall performance effectiveness and that they primarily rely on this perception when providing dimension ratings. Furthermore, we tested all previously described conventional models (i.e., the CS-, CD-, and CDCS-model) with an additional first-order general performance factor (cf. Hoffman et al., 2010; Hoffman et al., 2011; Lance, Lambert et al., 2004; Lance et al., 2000; Scullen, Mount, & Goff, 2000). For example, the CS1G-model hypothesized that although raters from different sources might capture different pieces of candidates' behavior, they have similar perceptions of candidates' overall performance effectiveness.

In the third set of tested models, dimensions were modeled by specifying broad dimension factors. That is, ratings of conceptually similar dimension were treated as manifest indicators of broad dimensions (cf. Hoffman et al., 2011). Recently, Hoffman et al. (2011) used this approach to evaluate within-exercise dimension ratings from different ACs and found consistent evidence for broad dimension factors. Comparable to Hoffman et al., we referred to common taxonomies of performance dimensions by Arthur et al. (2003), Borman and Brush (1993), and Shore et al. (1990) to classify dimensions into broad dimensions (see Tables A5 to A7 in the Appendix).

To determine whether a model adequately represented the latent factor structure of the data, we first determined whether the models converged to a proper solution. Models with inadmissible solutions or estimation problems were considered as being inappropriate and, therefore, were not further evaluated. Then, we evaluated the goodness-of-fit of models that converged to a proper solution. Referring to Hu and Bentler (1999), we used the root mean

square error of approximation (RMSEA), the standardized root mean squared residual (SRMR), the Comparative Fit Index (CFI), and the Tucker Lewis Index (TLI), whereby cut-off values of ≤ .06 for RMSEA, ≤ .08 for SRMR, and ≥ .95 for CFI and TLI indicate a good fit of the model.

Sample 1. The matrix with correlations among dimension ratings from different sources is presented in Table 1. The mean same dimension-different source correlation between overall dimension ratings from the AC and dimension ratings from the supervisory and customer assessment was $r = .21$. However, the mean different dimension-same source correlation within the AC was $r = .43$ and thus larger than the mean same dimension-different source correlation, which is problematic with regard to construct-related validity.

In the CFAs, only two models produced an admissible solution, namely the model with source factors only (CS-model) and the model with only a general performance factor (1G-model). Neither model yielded an acceptable fit to the data, but the CS-model was closer to an acceptable fit than the 1G-model (see Table 4). None of the models with dimension factors or with broad dimension factors converged to an admissible solution.

Table 1

Sample 1 – Means, Standard Deviations, and Correlations Between Overall Dimension Ratings From the AC and Ratings of the Same Dimensions From External Sources

Source/Dimensions	*M*	*SD*	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.	17.
Assessment center																			
1. Oral presentation	4.25	1.53																	
2. Written communication	4.26	1.46	.50**																
3. Interpersonal skills	4.30	1.51	.47**	.49**															
4. Planning and organizing	4.27	1.51	.45**	.34**	.51**														
5. Decision making	4.39	1.40	.38**	.41**	.48**	.38**													
6. Leadership	4.28	1.44	.40**	.38**	.45**	.40**	.45**												
Supervisory assessment																			
7. Oral presentation	4.58	1.45	.29**	.15**	.29**	.27**	.19**	.22**											
8. Written communication	4.43	1.42	.13**	.12*	.16**	.08	.14**	.12*	.46**										
9. Interpersonal skills	4.35	1.39	.19**	.13**	.29**	.23**	.18**	.16**	.48**	.47**									
10. Planning and organizing	4.55	1.42	.25**	.26**	.30**	.15**	.17**	.27**	.50**	.45**	.41**								
11. Decision making	4.80	1.40	.22**	.14**	.20**	.16**	.14**	.20**	.45**	.50**	.40**	.47**							
12. Leadership	4.89	1.37	.24**	.18**	.19**	.16**	.19**	.33**	.47**	.43**	.40**	.45**	.47**						
Customer assessment																			
13. Oral presentation	4.31	0.97	.21**	.17**	.26**	.25**	.28**	.23**	.64**	.34**	.35**	.37**	.30**	.38**					
14. Written communication	4.09	0.82	.16**	.11*	.26**	.17**	.20**	.18**	.36**	.56**	.31**	.33**	.30**	.28**	.37**				
15. Interpersonal skills	4.09	0.87	.21**	.15**	.31**	.22**	.27**	.21**	.31**	.32**	.55**	.27**	.25**	.29**	.36**	.33**			
16. Planning and organizing	4.18	0.99	.15**	.20**	.18**	.17**	.13*	.07	.37**	.35**	.27**	.43**	.32**	.32**	.37**	.38**	.25**		
17. Decision making	4.32	1.08	.17**	.17**	.19**	.12*	.22**	.10	.43**	.37**	.24**	.30**	.30**	.29**	.57**	.51**	.31**	.37**	
18. Leadership	4.34	1.02	.18**	.12*	.17**	.10*	.16**	.16**	.34**	.38**	.26**	.31**	.30**	.43**	.36**	.47**	.26**	.35**	.47**

Note. $N = 390$. * $p < .05$, ** $p < .01$ (two-tailed).

Sample 2. Correlations between dimension ratings from different sources are presented in Table 2. The mean same dimension-different source correlation between overall dimension ratings from the AC and external dimension ratings was $r = .11$, and the mean different dimension-same source correlation within the AC was $r = .30$. These results indicate that the AC did not have construct-related validity.

The CFAs yielded admissible solutions for three models (see Table 4): The model with source factors only (CS-model), the model with only a general performance factor (1G-model), and the model with two broad dimensions (2Bd-model). The model fit was poor for all three solutions, but in a comparative sense, the CS-model represented the data best. Models with conventional dimension factors did not converge to admissible solutions.

Table 2

Sample 2 – Means, Standard Deviations, and Correlations Between Overall Dimension Ratings From the AC and Ratings of the Same Dimensions From External Sources

Source/Dimensions	*M*	*SD*	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.	17.	18.
Assessment center																				
1. Achievement motivation	2.96	0.25	*(.24)*																	
2. Analysis	2.76	0.39	.40**	*(.30)*																
3. Interpersonal skills	2.97	0.23	.27**	.20*	*(.12)*															
4. Oral communication	3.02	0.21	.30**	.40**	.15	*(.49)*														
5. Dealing with conflicts	2.80	0.32	.41**	.20*	.31**	.14	*(.27)*													
6. Influencing others	2.72	0.45	.38**	.04	.32**	.36**	.53**	*(.42)*												
Supervisory assessment																				
7. Achievement motivation	3.55	0.96	.16	.09	-.11	.02	.05	.05	*(.97)*											
8. Analysis	3.28	0.96	.11	.19*	-.08	-.03	.10	.07	.61**	*(.97)*										
9. Interpersonal skills	3.34	0.88	.09	.07	-.00	.10	-.02	-.08	.33**	.43**	*(.96)*									
10. Oral communication	3.19	1.00	.18*	.21*	-.06	.16	.09	.06	.53**	.75**	.66**	*(.96)*								
11. Dealing with conflicts	3.16	0.89	-.02	.11	-.01	.02	.03	-.01	.45**	.60**	.64**	.69**	*(.95)*							
12. Influencing others	3.14	1.01	.21*	.19*	-.00	.07	.16	.08	.65**	.75**	.58**	.81**	.63**	*(.97)*						
Self-evaluation																				
13. Achievement motivation	3.71	0.78	-.09	-.06	.08	-.17	.00	-.04	.23*	.18	-.02	-.02	.07	.13	*(.90)*					
14. Analysis	3.72	0.54	.12	.15	-.05	-.02	.18	.02	.05	.31**	.02	.13	.03	.23*	.38**	*(.76)*				
15. Interpersonal skills	3.59	0.71	.01	.02	.02	.10	-.08	-.10	-.04	-.02	.26**	.09	.13	.05	.09	.08	*(.85)*			
16. Oral communication	3.65	0.50	.22*	.12	.06	.22*	.20*	.24**	.09	.15	.19*	.18*	.05	.20*	.15	.34**	.25**	*(.64)*		
17. Dealing with conflicts	3.48	0.63	.18*	.08	.15	.07	.17	.16	-.09	.15	.17	.17	.16	.16	.08	.35**	.24**	.29**	*(.83)*	
18. Influencing others	3.61	0.58	.19*	-.00	.18*	-.05	.17	.21*	.05	.09	.09	.05	.07	.24**	.30**	.25**	.20*	.41**	.27**	*(.80)*

Note. $N = 121$. * $p < .05$, ** $p < .01$ (two-tailed). Cronbach's α is reported in parentheses.

Sample 3. Table 3 shows correlations between dimension ratings from the AC and external sources. The mean same dimension-different source correlation between overall dimension ratings from the AC and external dimension ratings was $r = .12$, and the mean different dimension-same source correlation within the AC was $r = .50$, indicating that ratings of specific dimensions did not converge across sources and that the AC overall dimension ratings did not discriminate between dimensions.

In the CFAs, the model with source factors only (CS-model), the model with only a general performance factor (1G-model), the model with two broad dimensions (2Bd-model), and the model with three broad dimensions (3Bd-model) converged to an admissible solution (see Table 4). All converging models generated a poor model fit, but the CS-model was closest to an acceptable fit. Models that contained conventional dimension factors did not yield admissible solutions.

As mentioned above, we found differences in the internal consistencies of the dimension evaluations from the AC. As a lack of reliability of dimension ratings might be a reason why construct-related validity cannot be established (cf. Arthur et al., 2008), we repeated the CFAs and used only dimensions with an acceptable internal consistency in the AC. That is, for the second set of CFAs, we only used organizing and planning, presentation skills, and persuasiveness. Thereby, four models converged to an admissible solution (see Table 4): The model with source factors only (CS-model), the model with only a general performance factor (1G-model), the model with a combination of source factors and a general performance factor (CS1G-model), and the model with two broad dimensions and a general performance factor (2Bd1G-model). Models with conventional dimension factors did not converge to admissible solutions.

Of all converging models, the different fit indices only indicated a good fit to the data for the CS-model and the CS1G-model. In the CS-model, source factors explained an average

of 61% of variance in dimension ratings, and in the CS1G-model, the respective values were 62% for source factors and 6% for the general performance factor.

All fit indices of the CS1G-model were slightly better than those of the CS-model. To determine which of these two models was more appropriately representing the latent factor structure of the data, we considered the $\Delta\chi^2$. Furthermore, we used two additional comparative indices, ΔCFI and the relative fit index (RFI; see, for example, Hoffman et al., 2010; Lakey, Goodie, Lance, Stinchfield, & Winters, 2007; Lance, Foster et al., 2004). For ΔCFI, cut-off values between .002 and .01 have been suggested as indicating a significant difference in the goodness-of-fits of two models (Cheung & Rensvold, 2002; Meade, Johnson, & Braddy, 2008). Concerning the RFI (cf. Equation 1), this index allows a comparison between the fit of a more restrictive model (M_R, in our case the CS-model) relative to the fit of a less restrictive model (M_U, in our case the CS1G-model) as compared to the null model (M_{Null}). RFI values range from 0 to 1, whereby values close to 1 indicate that the two models are comparable with regard to their goodness-of-fit.

$$RFI = 1 - \frac{\chi^2_{M_R} - \chi^2_{M_U}}{\chi^2_{Null} - \chi^2_{M_U}} \qquad \text{(Equation 1)}$$

The ΔCFI value of .019 indicated that the goodness-of-fit of the CS1G-model was better relative to the CS-model, irrespective of which cutoff value we referred to. In contrast, the $\Delta\chi^2$ test, $\Delta\chi^2(9) = 15.09$, $p = .09$, and the RFI value of .96 indicated that the CS1G-model and the CS-model were statistically equivalent. Thus, from a practical standpoint, the CS-model seems to explain the data sufficiently well, so that no additional general performance factor is needed.

Table 3

Sample 3 – Means, Standard Deviations, and Correlations Between Overall Dimension Ratings From the AC and Ratings of the Same Dimensions From External Sources

Source/Dimensions	*M*	*SD*	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.	17.	18.
Assessment center																				
1. Analytical skills	3.17	0.72	*(.35)*																	
2. Organizing and planning	3.19	0.73	.56**	*(.76)*																
3. Persuasiveness	3.41	0.59	.65**	.74**	*(.69)*															
4. Assertiveness	3.00	0.88	.42**	.60**	.70**	*(.66)*														
5. Cooperation	3.14	0.89	.34**	.47**	.35**	.29**	-													
6. Presentation skills	3.42	0.84	.47**	.63**	.57**	.31**	.19	*(.71)*												
Supervisory assessment																				
7. Analytical skills	4.13	0.64	.18	.32**	.17	.15	.19	.25*	*(.86)*											
8. Organizing and planning	4.24	0.65	.14	.16	.04	.13	.06	.09	.72**	*(.82)*										
9. Persuasiveness	4.02	.0.64	.14	.32**	.15	.24*	.10	.28**	.71**	.57**	*(.85)*									
10. Assertiveness	3.90	0.64	.02	.22*	.11	.23*	-.01	.22*	.60**	.53**	.79**	*(.75)*								
11. Cooperation	4.14	0.56	-.10	-.02	-.15	-.10	.01	-.08	.37**	.25*	.25*	.29**	*(.70)*							
12. Presentation skills	4.09	0.60	.06	.25*	.02	.11	.11	.18	.55**	.48**	.68**	.62**	.33**	*(.71)*						
Self-evaluation																				
13. Analytical skills	4.06	0.54	.03	.11	.13	.09	-.05	.17	.23*	.28**	.26*	.23*	.13	.21*	*(.78)*					
14. Organizing and planning	4.08	0.58	-.03	.10	.08	.02	.02	.03	.21*	.35**	.22*	.23*	.04	.22*	.70**	*(.85)*				
15. Persuasiveness	4.02	0.62	.00	.23*	.12	.13	.04	.18	.25*	.26*	.37**	.27*	.15	.32**	.67**	.60**	*(.89)*			
16. Assertiveness	3.91	0.62	-.14	.16	.06	.13	.04	.05	.24*	.20	.36**	.27**	.06	.22*	.55**	.51**	.79**	*(.83)*		
17. Cooperation	3.86	0.49	-.11	-.05	-.08	-.08	.03	.03	.20	.27**	.15	.16	.16	.20	.45**	.50**	.35**	.15	*(.74)*	
18. Presentation skills	4.22	0.51	-.14	.19	.03	.06	.00	.15	.06	.11	.24*	.23*	.15	.22*	.53**	.51**	.64**	.64**	.28**	*(.75)*

Note. $N = 92$. * $p < .05$, ** $p < .01$ (two-tailed). Cronbach's α is reported in parentheses. Cooperation was rated in one exercise only, therefore, no Cronbach's α is reported in this case.

Table 4

Model Fit Statistics for the Structure of Overall Dimension Ratings From the AC and Dimension Ratings From External Sources for Models That Converged to a Proper Solution

Sample and model	*df*	χ^2	RMSEA	SRMR	TLI	CFI
Sample 1						
Conventional models						
CS	132	527.95**	.084	.057	.838	.860
Conventional models with a general performance factor						
1G	135	1099.20**	.129	.106	.615	.660
Sample 2						
Conventional models						
CS	132	268.54**	.093	.082	.796	.824
Conventional models with a general performance factor						
1G	135	444.89**	.138	.137	.547	.600
Models with broad dimensions						
2Bd	134	436.46**	.137	.137	.554	.610
Sample 3 (all dimensions used for analyses)						
Conventional models						
CS	132	223.76**	.087	.076	.877	.894
Conventional models with a general performance factor						
1G	135	655.26**	.206	.191	.318	.398
Models with broad dimensions						
2Bd	134	642.30**	.204	.197	.328	.412
3Bd	132	640.98**	.206	.198	.317	.411
Sample 3 (only dimensions with an acceptable internal consistency used for analyses)						
Conventional models						
CS	24	30.19	.053	.056	.972	.981
Conventional models with a general performance factor						
1G	27	210.03**	.273	.180	.251	.438
CS1G	15	15.09	.008	.043	.999	1.00
Models with broad dimensions						
2Bd1G	17	90.85**	.218	.110	.520	.773

Note. Sample sizes were $N = 390$ for Sample 1, $N = 121$ for Sample 2, and $N = 92$ for Sample 3. In Sample 3, dimensions with acceptable internal consistency were organizing and planning, presentation skills, and persuasiveness. CD = correlated dimensions, CS = correlated sources, Bd = broad dimension, G = general performance factor. ** $p < .01$.

Discussion

We examined the external construct-related validity of ACs by relating overall dimension ratings from an AC to comparison scores provided from sources external to the AC as has been repeatedly proposed (cf. Arthur et al., 2008; Neidig & Neidig, 1984; Reilly et al., 1990; Rupp et al., 2008). Contrary to prior research, the external comparison scores referred to the same dimensions as the AC overall dimension ratings, meaning that constructs were held constant across methods. The AC overall dimension ratings were also independent from external comparison scores. This is in contrast to previous studies that, for example, incorporated peer- or self-evaluations of candidates' performance in the AC into AC overall dimension ratings (e.g., Shore et al., 1992). These methodological strengths allowed us to clearly separate method effects from dimension effects and to provide an answer to the question of how AC overall dimension ratings and external ratings of the same dimensions are related to each other (cf. Arthur et al., 2008; Neidig & Neidig, 1984; Reilly et al., 1990; Rupp et al., 2008). The use of three different samples (two samples from field settings and one sample from a laboratory setting) enabled us to draw firm conclusions on the generalizability of the results obtained.

Our study contributes to the literature in several respects. First, our results demonstrate that relating AC overall dimension ratings to external evaluations of identical dimensions does not provide evidence for AC construct-related validity as expected by some researchers (cf. Arthur et al., 2008; Neidig & Neidig, 1984; Reilly et al., 1990; Rupp et al., 2008). Additionally, our study suggests that a lack of construct-related validity is not primarily due to unreliability of ratings (cf. Arthur et al., 2008; Howard, 2008). Also, when integrating multiple dimension ratings into overall dimension ratings, that is, when increasing the number of "items" that constitute a dimension rating, construct-related validity will not necessarily be

established. This is also true when overall dimension ratings reach an acceptable level of internal consistency (see Sample 3).

Second, our study revealed that ACs offer a perspective on employees' performance that differs substantially from those of other sources. Specifically, ACs allow observing performance in a selection context, whereas other sources provide information on performance on the job. This implies that only using information from both an AC and from other sources allows for a comprehensive picture of employees' performance. That is, ACs cannot be replaced by multisource feedback, for example, when making placement decisions or decisions concerning employees' developmental needs.

Third, based on the aforementioned contributions and as outlined below, our study offers some practical guidance concerning the use of dimension ratings when providing feedback to candidates, trying to identify employees' developmental needs, or determining whether a candidate's profile suits the demands of a particular position.

Concerning the contributions of our results to the literature, a consistent finding across all three samples was that evidence for the external construct-related validity of the ACs was poor on the correlational level as well as with regard to the latent factor structure of AC overall dimension ratings and external ratings of the same dimensions. Different dimension-same source correlations within the ACs were larger than same dimension-different source correlations. In line with this, CFAs revealed source factors or a general performance factor in all three datasets. Models with conventional dimension factors did not converge to a proper solution in any of the samples, but in two samples, models with broad dimension factors also converged to a proper solution. However, goodness-of-fit statistics indicated that, in general, models with source factors represented the factor structure underlying AC overall dimension ratings and external dimension ratings best. Thus, in models that incorporated AC overall dimension ratings and external ratings of the same dimensions, dimension factors did not

seem to be an important source of variance. Furthermore, CFA results were similar when only dimensions that reached an acceptable level of internal consistency were used for analyses (as in Sample 3). In this case, however, the model with a combination of source factors and a general performance factor and the model with only source factors were similarly appropriate for the latent factor structure of AC overall dimension ratings and external dimension ratings. Yet, two of the three comparative indices used indicated that source factors alone sufficed to explain the variance in the data. Furthermore, compared to source factors, the general performance factor accounted for only a small amount of explained variance in ratings. Thus, a general performance factor was not necessarily needed in the latent factor structure of AC overall dimension ratings and external dimension ratings.

The results concerning the correlations between AC overall dimension ratings and external dimension ratings and the absence of dimension factors in the latent factor structure of AC overall dimension ratings and external dimension ratings suggests that AC overall dimension ratings cannot be attributed to the targeted dimensions. This finding does not support the suggestion that AC construct-related validity can be evidenced when using AC overall dimension ratings as focal constructs for validation in combination with dimension evaluations that stem from other sources (cf. Arthur et al., 2008; Neidig & Neidig, 1984; Reilly et al., 1990; Rupp et al., 2008). Rather, our findings conform with arguments against this expectation that can be inferred from theoretical assumptions of multisource feedback (Borman, 1974; 1977) and related empirical findings (Conway, 1996; Hoffman et al., 2010; Woehr et al., 2005). The predominance of source variance in ratings from different sources implies that candidates in general behave differently in interactions with different sources, which is in line with interactionist theories (Mischel & Shoda, 1995; Tett & Burnett, 2003; Tett & Guterman, 2000). As a consequence of the situational specificity of candidates' behavior, different sources observe different behaviors. Furthermore, these sources might

capture different slices of behavior and, therefore, provide different ratings (cf. Borman, 1974; 1977). Taken together, different sources seem to offer different perspectives on candidates' overall performance. Thus, the AC and other sources all provide important information for the evaluation of candidates' performance.

It seems unlikely that the findings of the present study can be attributed to special characteristics of the ACs considered: All three ACs were comparable to other ACs found in the literature and in the field with respect to design characteristics like, for example, the kind and number of dimensions used in the AC, the number of observed dimensions per exercise, the types of exercises, and assessor training (cf. Eurich, Krause, Cigularov, & Thornton, 2009; Krause & Thornton, 2009; Woehr & Arthur, 2003). However, as mentioned above, the ACs were also comparable to other ACs concerning their internal construct-related validity (e.g., Melchers et al., 2007; Woehr & Arthur, 2003) and their criterion-related validity (e.g., Becker et al., 2011; Gaugler et al., 1987; Hardison & Sackett, 2004; Hermelin et al., 2007). Furthermore, we did not even find dimension factors in the latent factor structure of dimension ratings when we used only AC overall dimension ratings with an acceptable internal consistency for the analyses, indicating also that the poor internal consistency of AC overall dimension ratings does not explain our results.

Our results might seem to be at odds with findings from previous studies that offered support for the external construct-related validity of ACs (e.g., Dilchert & Ones, 2009; Shore et al., 1992; Shore et al., 1990; Thornton et al., 1997). A possible reason for the diverging results from the present study is that we related AC overall dimension ratings to performance evaluations that referred to the job context. However, as mentioned above, the AC captures other aspects of performance than can be observed on the job. Therefore, convergence between AC dimension ratings and dimension ratings from external sources might be low. Contrary to our study, previous studies related AC ratings to other variables that were also

gathered in a selection context like, for example, cognitive ability measures or peers' evaluations of candidates' AC performance (e.g., Shore et al., 1992; Shore et al., 1990). Thus, those previous studies related AC ratings to other variables obtained in situations in which people were motivated to perform at peak level, which might have increased the probability of convergence (Ployhart et al., 2001). Therefore, the probability of finding dimension factors and thus evidence for external construct-related validity of AC when relating AC overall dimension ratings to external ratings of the same dimensions should increase when using dimension ratings from other maximum performance situations as comparative data to AC overall dimension ratings. Future research is needed to test this possibility.

Finally, it needs to be mentioned that a general performance factor has repeatedly been found in the internal structure of AC ratings (Hoffman et al., 2011; Lance, Foster et al., 2004; Lance, Foster, Nemeth, Gentry, & Drollinger, 2007; Lance et al., 2000). At first glance, this finding seems to be inconsistent with our results. However, when examining the internal structure of AC ratings, the focus lies on dimension ratings from different exercises. As single elements of an AC, different exercises all represent maximum performance situations. Therefore, when examining the internal structure of within-exercise dimension ratings, different evaluations of maximum performance are related to each other and thus finding a general performance factor in the latent factor structure of these ratings is not surprising. It is possible that we would also have found a general performance factor in the latent factor structure of AC overall dimension ratings and external dimension ratings if we had used comparative data from other situations that evoke maximum performance, for example, an interview.

Despite the lack of evidence for the external construct-related validity of ACs reported in the present study, the ACs considered were criterion valid. In both samples in which we were able to determine AC criterion-related validity (Samples 2 and 3), the overall

performance in the ACs was significantly related to performance criteria evaluated by participants' supervisors. These results are in line with previous findings on AC criterion-related validity (e.g., Becker et al., 2011; Gaugler et al., 1987; Hardison & Sackett, 2004; Hermelin et al., 2007) and indicate that the ACs were comparable to other ACs found in the literature in this regard. Hence, it seems that requirements of the job were well represented by the exercises and, therefore, participants' behavior in these exercises allowed for a prediction of job performance. However, the ACs used in Sample 2 and Sample 3 differed with regard to criterion-related validity. This difference in criterion-related validity of these two ACs might reflect differences in the representativeness of the ACs for the job context. In Sample 2, the AC was designed for the selection of prospective career officers, and criterion measures referred to the performance in career officer training. In Sample 3, we used a graduate AC that covered a wide range of requirements essential for a variety of jobs. However, due to the fact that participants' jobs for which criterion data were obtained were very heterogeneous, the AC was differently representative for different jobs. This possibly led to a lower criterion-related validity of the AC used in Sample 3 compared to the AC used in Sample 2.

Practical Implications

As AC overall dimension ratings do not seem to be attributable to dimensions, it might be recommended to focus more on specific aspects of AC performance that are related to job performance than on dimension ratings when making placement decisions and decisions concerning employees' developmental needs, or when providing feedback to candidates concerning their AC performance. In light of findings on the internal structure of AC ratings, referring to the performance in specific exercises (cf. Lance, Lambert et al., 2004) would be an option.

Our findings are of importance not only for the AC domain but also for multisource feedback. It can be inferred from past research that multisource feedback might be a substitute for ACs (e.g., Hagan et al., 2006). However, our results suggest that the AC and other sources that refer to the job context capture different aspects of performance. Therefore, ACs and multisource feedback should be regarded as different methods that provide different perspectives on candidates' performance and not as mere substitutes for each other. From a practical point of view, using information from an AC and other sources (but not aggregating this information across sources) might allow for a more comprehensive picture of candidates' strengths and weaknesses than a sole AC or multisource feedback program. This might be useful especially for developmental purposes and placement decisions. However, when performance evaluations were obtained from an AC and other sources, feedback to candidates should be source-specific. Thereby, addressing differences in the perception of candidates' performance between sources and also differences between typical and maximum performance might provide valuable information to candidates.

Limitations and Suggestions for Future Research

Several limitations of the present study should be noted. First, in Sample 1, dimension ratings from the AC and the external sources were one-item measures. The reliability of these ratings might have been improved if multiple items for each dimension were used.

Second, especially in Sample 1, the difficulty of evaluating some of the dimensions of interest might have differed across sources, which might have influenced our results (Murphy & Cleveland, 1995). For example, it might have been more difficult to provide ratings on the dimension of leadership for customers than for supervisors. Therefore, future research that aims to compare methods should ensure that all sources or methods, respectively, are equally capable of evaluating the focal constructs.

The third limitation concerns Sample 3. The AC used in Sample 3 was designed to cover requirements that are essential in many graduate jobs. However, due to the heterogeneity of the participants' jobs, we assume that in some cases the exercises represented the requirements of the jobs better than in other cases. Furthermore, the AC dimensions were probably of varying importance for participants' jobs. On the one hand, these differences in the representativeness of the AC for participants' jobs might have reduced AC criterion-related validity. On the other hand, they might have led to differences in the degree to which AC overall dimension ratings and external dimension ratings converged and thus might have contributed to the fact that no dimension factors were found.

Fourth, we compared AC overall dimension ratings to external dimension ratings that referred to the job context. As already mentioned, the AC is assumed to evoke maximum performance, whereas on the job usually only typical performance can be observed. This might have reduced the probability of convergence between the different sources and of finding common underlying factors in the latent structure of AC overall dimension ratings and dimension ratings provided by other sources. Therefore, future research might relate AC overall dimension ratings to evaluations of candidates' performance in other maximum performance situations to examine whether dimension factors can be found for AC overall dimension ratings and external dimension ratings. Potential situations that could be used for comparison to the AC are assessment situations that allow evaluating the same dimensions as used in the focal AC, that is, a parallel AC or an interview, for example.

Conclusion

In the present study, consistent findings across three samples lead to the conclusion that AC overall dimension ratings and ratings of the same dimensions provided from other sources cannot be attributed to dimension factors. We did not find dimension factors in the

latent factor structure of AC overall dimension ratings and external dimension ratings that referred to the job context when following recent developments and promising findings in the AC domain by modeling broad dimension factors (e.g., Hoffman et al., 2011). Furthermore, our results did not support a common general performance factor for dimension ratings from the AC and from external sources. Our findings suggest that the AC provides a different perspective on people's performance than other sources and that different aspects of performance are captured in the AC than in the job context. However, despite the lack of evidence for dimension factors in the latent factor structure of AC overall dimension ratings and external dimension ratings, we found support for AC criterion-related validity, indicating that the ACs measured something that was critical to job performance. Therefore, and in light of support for the incremental validity of AC performance beyond cognitive ability or personality (e.g., Dilchert & Ones, 2009; Krause et al., 2006; Melchers & Annen, 2010; Meriac et al., 2008), we are still convinced that ACs can be an important source of information for personnel decisions.

References

Arthur, W., Jr., Day, E. A., McNelly, T. L., & Edens, P. S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology, 56*, 125-154. doi:10.1111/j.1744-6570.2003.tb00146.x

Arthur, W., Jr., Day, E. A., & Woehr, D. J. (2008). Mend it, don't end it: An alternate view of assessment center construct-related validity evidence. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 105-111. doi:10.1111/j.1754-9434.2007.00019.x

Arthur, W., Jr., & Villado, A. J. (2008). The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. *Journal of Applied Psychology, 93*, 435-442. doi:10.1037/0021-9010.93.2.435

Atkins, P. W., & Wood, R. E. (2002). Self- versus others' ratings as predictors of assessment center ratings: Validation evidence for 360-degree feedback programs. *Personnel Psychology, 55*, 871-904. doi:10.1111/j.1744-6570.2002.tb00133.x

Becker, N., Höft, S., Holzenkamp, M., & Spinath, F. M. (2011). The predictive validity of assessment centers in German-speaking regions: A meta-analysis. *Journal of Personnel Psychology, 10*, 61-69. doi:10.1027/1866-5888/a000031

Borman, W. C. (1974). The rating of individuals in organizations: An alternate approach. *Organizational Behavior & Human Performance, 12*, 105-124. doi:10.1016/0030-5073(74)90040-3

Borman, W. C. (1997). 360° ratings: An analysis of assumptions and a research agenda for evaluating their validity. *Human Resource Management Review, 7*, 299-315. doi:10.1016/S1053-4822(97)90010-3

Borman, W. C., & Brush, D. H. (1993). More progress toward a taxonomy of managerial performance requirements. *Human Performance, 6*, 1-21. doi:10.1207/s15327043hup0601_1

Bowler, M. C., & Woehr, D. J. (2006). A meta-analytic evaluation of the impact of dimension and exercise factors on assessment center ratings. *Journal of Applied Psychology, 91*, 1114-1124. doi:10.1037/0021-9010.91.5.1114

Chan, D. (1996). Criterion and construct validation of an assessment centre. *Journal of Occupational and Organizational Psychology, 69*, 167-181. doi:10.1111/j.2044-8325.1996.tb00608.x

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233-255. doi:10.1207/S15328007SEM0902_5

Conway, J. M. (1996). Analysis and design of multitrait-multirater performance appraisal studies. *Journal of Management, 22*, 139-162. doi:10.1177/014920639602200106

Conway, J. M., & Huffcutt, A. I. (1997). Psychometric properties of multisource performance ratings: A meta-analysis of subordinate, supervisor, peer, and self-ratings. *Human Performance, 10*, 331-360. doi:10.1207/s15327043hup1004_2

Dilchert, S., & Ones, D. S. (2009). Assessment center dimensions: Individual differences correlates and meta-analytic incremental validity. *International Journal of Selection and Assessment, 17*, 254-270. doi:10.1111/j.1468-2389.2009.00468.x

Eurich, T. L., Krause, D. E., Cigularov, K., & Thornton, G. C., III. (2009). Assessment centers: Current practices in the United States. *Journal of Business and Psychology, 24*, 387-407. doi:10.1007/s10869-009-9123-3

Fleenor, J. W. (1996). Constructs and developmental assessment centers: Further troubling empirical findings. *Journal of Business and Psychology, 10*, 319-335. doi:10.1007/BF02249606

Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., III, & Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology, 72*, 493-511. doi:10.1037/0021-9010.72.3.493

Goffin, R. D., Jelley, R. B., Powell, D. M., & Johnston, N. G. (2009). Taking advantage of social comparisons in performance appraisal: The relative percentile method. *Human Resource Management, 48*, 251-268. doi:10.1002/hrm.20278

Gutknecht, S. P., Semmer, N. K., & Annen, H. (2005). Prognostische Validitat eines Assessment Centers fur den Studien- und Berufserfolg von Berufsoffizieren der Schweizer Armee / Predictive validity of an assessment center for the study and professional success of career officers in the Swiss Army. *Zeitschrift für Personalpsychologie, 4*, 170-180. doi:10.1026/1617-6391.4.4.170

Hagan, C. M., Konopaske, R., & Bernardin, H. J. (2006). Predicting assessment center performance with 360-degree, top-down, and customer-based competency assessments. *Human Resource Management, 45*, 357-390. doi:10.1002/hrm.20117

Hardison, C. M., & Sackett, P. R. (2004). *Assessment center criterion related validity: A meta-analytic update*. Unpublished manuscript.

Hermelin, E., Lievens, F., & Robertson, I. T. (2007). The validity of assessment centres for the prediction of supervisory performance ratings: A meta-analysis. *International Journal of Selection and Assessment, 15*, 405-411. doi:10.1111/j.1468-2389.2007.00399.x

Hoffman, B., Lance, C. E., Bynum, B., & Gentry, W. A. (2010). Rater source effects are alive and well after all. *Personnel Psychology, 63*, 119-151. doi:10.1111/j.1744-6570.2009.01164.x

Hoffman, B., Melchers, K. G., Blair, C. A., Kleinmann, M., & Ladd, R. T. (2011). Exercises AND dimensions are the currency of assessment centers. *Personnel Psychology, 64*, 351-395. doi:10.1111/j.1744-6570.2011.01213.x

Howard, A. (1997). A reassessment of assessment centers: Challenges for the 21st century. *Journal of Social Behavior and Personality, 12*, 13-52.

Howard, A. (2008). Making assessment centers work the way they are supposed to. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 98-104. doi:10.1111/j.1754-9434.2007.00018.x

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55. doi:10.1080/10705519909540118

Kolk, N. J., Born, M. P., & van der Flier, H. (2002). Impact of common rater variance on construct validity of assessment center dimension judgments. *Human Performance, 15*, 325-338. doi:10.1207/S15327043HUP1504_02

Krause, D. E., Kersting, M., Heggestad, E. D., & Thornton, G. C., III. (2006). Incremental validity of assessment center ratings over cognitive ability tests: A study at the executive management level. *International Journal of Selection and Assessment, 14*, 360-371. doi:10.1111/j.1468-2389.2006.00357.x

Krause, D. E., & Thornton, G. C., III. (2009). A cross-cultural look at assessment center practices: Survey results from Western Europe and North America. *Applied Psychology: An International Review, 58*, 557-585. doi:10.1111/j.1464-0597.2008.00371.x

Lakey, C. E., Goodie, A. S., Lance, C. E., Stinchfield, R., & Winters, K. C. (2007). Examining DSM-IV criteria for pathological gamling: Psychomeric properties and evidence from cognitive biases. *Journal of Gambling Studies, 23*. doi:10.1007/s10899-007-9063-7

Lance, C. E., Foster, M. R., Gentry, W. A., & Thoresen, J. D. (2004). Assessor cognitive processes in an operational assessment center. *Journal of Applied Psychology, 89*, 22-35. doi:10.1037/0021-9010.89.1.22

Lance, C. E., Foster, M. R., Nemeth, Y. M., Gentry, W. A., & Drollinger, S. (2007). Extending the nomological network of assessment center construct validity: Prediction of cross-situationally consistent and specific aspects of assessment center performance. *Human Performance, 20*, 345-362.

Lance, C. E., Lambert, T. A., Gewin, A. G., Lievens, F., & Conway, J. M. (2004). Revised estimates of dimension and exercise variance components in assessment center postexercise dimension ratings. *Journal of Applied Psychology, 89*, 377-385. doi:10.1037/0021-9010.89.2.377

Lance, C. E., Newbolt, W. H., Gatewood, R. D., Foster, M. R., French, N. R., & Smith, D. E. (2000). Assessment center exercise factors represent cross-situational specificity, not method bias. *Human Performance, 13*, 323-353. doi:10.1207/S15327043HUP1304_1

Lievens, F., & Conway, J. M. (2001). Dimension and exercise variance in assessment center scores: A large-scale evaluation of multitrait-multimethod studies. *Journal of Applied Psychology, 86*, 1202-1222. doi:10.1037/0021-9010.86.6.1202

Lievens, F., Dilchert, S., & Ones, D. S. (2009). The importance of exercise and dimension factors in assessment centers: Simultaneous examinations of construct-related and criterion-related validity. *Human Performance, 22*, 375-390. doi:10.1080/08959280903248310

Lievens, F., Tett, R. P., & Schleicher, D. J. (2009). Assessment centers at the crossroads: Toward a reconceptualization of assessment center exercises. In J. J. Martocchio & H. Liao (Eds.), *Research in Personnel and Human Resources Management* (pp. 99-152). Bingley: JAI Press.

Marcus, B., Goffin, R. D., Johnston, N. G., & Rothstein, M. G. (2007). Personality and cognitive ability as predictors of typical and maximum managerial performance. *Human Performance, 20*, 275-285. doi:10.1080/08959280701333362

McCloy, R. A., Campbell, J. P., & Cudeck, R. (1994). A confirmatory test of a model of performance determinants. *Journal of Applied Psychology, 79*, 493-505. doi:10.1037/0021-9010.79.4.493

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology, 93*, 568-592. doi:10.1037/0021-9010.93.3.568

Melchers, K. G., & Annen, H. (2010). Officer selection for the Swiss armed forces: An evaluation of validity and fairness issues. *Swiss Journal of Psychology, 69*, 105-115. doi:10.1024/1421-0185/a000012

Melchers, K. G., Henggeler, C., & Kleinmann, M. (2007). Do within-dimension ratings in assessment centers really lead to improved construct validity? A meta-analytic reassessment. *Zeitschrift für Personalpsychologie, 6*, 141-149. doi:10.1026/1617-6391.6.4.141

Melchers, K. G., Wirz, A., & Kleinmann, M. (in press). Dimensions AND exercises: Theoretical background of mixed-model assessment centers. In D. J. R. Jackson, C. E. Lance & B. J. Hoffman (Eds.), *The psychology of assessment centers*. New York: Routledge.

Meriac, J. P., Hoffman, B. J., Woehr, D. J., & Fleisher, M. S. (2008). Further evidence for the validity of assessment center dimensions: A meta-analysis of the incremental criterion-related validity of dimension ratings. *Journal of Applied Psychology, 93*, 1042-1052. doi:10.1037/0021-9010.93.5.1042

Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review, 102*, 246-268. doi:10.1037/0033-295X.102.2.246

Murphy, K. R., & Cleveland, J. N. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives*. Thousand Oaks, CA: Sage.

Neidig, R. D., & Neidig, P. J. (1984). Multiple assessment center exercises and job relatedness. *Journal of Applied Psychology, 69*, 182-186. doi:10.1037/0021-9010.69.1.182

Ployhart, R. E., Lim, B. C., & Chan, K. Y. (2001). Exploring relations between typical and maximum performance ratings and the five factor model of personality. *Personnel Psychology, 54*, 809-843. doi:10.1111/j.1744-6570.2001.tb00233.x

Reilly, R. R., Henry, S., & Smither, J. W. (1990). An examination of the effects of using behavior checklists on the construct validity of assessment center dimensions. *Personnel Psychology, 43*, 71-84. doi:10.1111/j.1744-6570.1990.tb02006.x

Rupp, D. E., Thornton, G. C., III, & Gibbons, A. M. (2008). The construct validity of the assessment center method and usefulness of dimensions as focal constructs. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 116-120. doi:10.1111/j.1754-9434.2007.00021.x

Sackett, P. R., Zedeck, S., & Fogli, L. (1988). Relations between measures of typical and maximum job performance. *Journal of Applied Psychology, 73*, 482-486. doi:10.1037/0021-9010.73.3.482

Scullen, S. E., Mount, M. K., & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology, 85*, 956-970. doi:10.1037/0021-9010.85.6.956

Shore, T. H., Shore, L. M., & Thornton, G. C., III. (1992). Construct validity of self- and peer evaluations of performance dimensions in an assessment center. *Journal of Applied Psychology, 77*, 42-54. doi:10.1037/0021-9010.77.1.42

Shore, T. H., Thornton, G. C., III, & Shore, L. M. (1990). Construct validity of two categories of assessment center dimension ratings. *Personnel Psychology, 43*, 101-116. doi:10.1111/j.1744-6570.1990.tb02008.x

Staufenbiel, T., & Hartz, C. (2000). Organizational citizenship behavior: Development and validation of a measurement instrument. *Diagnostica, 46*, 73-83. doi:10.1026//0012-1924.46.2.73

Tett, R. P., & Burnett, D. D. (2003). A personality trait-based interactionist model of job performance. *Journal of Applied Psychology, 88*, 500-517. doi:10.1037/0021-9010.88.3.500

Tett, R. P., & Guterman, H. A. (2000). Situation trait relevance, trait expression, and cross-situational consistency: Testing a principle of trait activation. *Journal of Research in Personality, 34*, 397-423. doi:10.1006/jrpe.2000.2292

Thornton, G. C., III, Tziner, A., Dahan, M., Clevenger, J. P., & Meir, E. (1997). Construct validity of assessment center judgements: Analyses of the behavioral reporting method. *Journal of Social Behavior and Personality, 12*, 109-128.

Woehr, D. J., & Arthur, W., Jr. (2003). The construct-related validity of assessment center ratings: A review and meta-analysis of the role of methodological factors. *Journal of Management, 29*, 231-258. doi:10.1177/014920630302900206

Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology, 67*, 189-205. doi:10.1111/j.2044-8325.1994.tb00562.x

Woehr, D. J., Sheehan, M., & Bennett, W., Jr. (2005). Assessing measurement equivalence across rating sources: A multitrait-multirater approach. *Journal of Applied Psychology, 90*, 592-600. doi:10.1037/0021-9010.90.3.592

Appendix

Table A1

Definitions of the AC Dimension Used in Sample 2

Dimension	Definition
Achievement motivation	Showing initiative, commitment, and persistence; accepting considerable strains and frustration to achieve ambitious goals.
Analysis	Taking on problems purposefully and in a structured manner; recognizing important connections, getting an overview, and forming a well-founded judgment through networked thinking; developing logical and flexible solutions to problems; setting priorities.
Interpersonal skills	Facing others with openness and fairness; being interested in and trying to understand needs of others; being able to empathize with others without giving up one's own position; being able to fit into a team and to cooperate.
Dealing with conflicts	Recognizing conflict potential; offering consensual solutions.
Oral communication	Being able to express oneself clearly; being able to listen actively; passing on a message with correspondence between the verbal and non-verbal; facing others directly.
Influencing others	Being able to convince others; motivating others; being able to influence another person's or a group's actions to promote the achievement of objectives.

Table A2

Dimension by Exercise Matrix in Sample 2

Dimension	SOP	LGD	MOT	DEB	SCS	LEC
Achievement motivation	X		X		X	X
Analysis	X				X	X
Interpersonal skills		X	X	X	X	
Dealing with conflicts		X	X		X	
Oral communication	X	X	X	X		X
Influencing others		X	X	X		

Note. SOP = short oral presentation, LGD = leaderless group discussion, MOT = motivational talk, DEB = debate, SCS = short case scenarios, LEC = lecture.

Table A3

Definitions of the AC Dimension Used in Sample 3

Dimension	Definition
Analytical skills	Analyzing carefully; quickly and correctly comprehending new contents; correctly recognizing connections; differentiating between important and unimportant.
Persuasiveness	Clearly explaining one's decisions; presenting solid arguments; selling one's ideas to others.
Organizing and planning	Being systematic; differentially organizing information; structuring presentations or discussions in a useful way; adequately estimating time requirements.
Assertiveness	Pushing one's interests even in light of resistance from others; not letting oneself get discouraged by others; acting in a determined way.
Cooperation	Picking up ideas that differ from one's own view; being willing to adapt one's view; helping to achieve objectives of the group.
Presentation skills	Appearing confident; speaking calmly and clearly; using gestures and mimic to support the verbal; turning toward listeners; maintaining eye contact with listeners.

Table A4

Dimension by Exercise Matrix in Sample 3

Dimension	SP	PLA	LGD ST	LGD GM	LGD WLB
Analytical skills	X		X		
Organizing and planning	X	X	X	X	X
Persuasiveness	X	X	X	X	X
Assertiveness				X	X
Cooperation			X		
Presentation skills	X	X			

Note. LGD = leaderless group discussion. SP = sales presentation, PLA = presentation of a leisure activity, LGD ST = staffing task, LGD GM = graduate marketing task, LGD WLB = work-life-balance task.

Table A5

Classification of the AC Dimensions Used in Sample 1 Into Broad Dimensions Based on Popular Taxonomies

Dimension	Arthur at al. (2003)	Borman & Brush (1993)	Shore et al. (1990)
Oral presentation	Communication	Interpersonal dealings and communication	Interpersonal style
Written communication			
Interpersonal skills	Consideration and awareness of others		
Leadership	Influencing others	Leadership	
Decision making	Problem solving	Technical activities and the mechanics of management	Performance style
Planning and organizing	Organizing and planning		

Table A6

Classification of the AC Dimensions Used in Sample 2 Into Broad Dimensions Based on Popular Taxonomies

<table>
<tr><th>Dimension</th><th>Arthur at al. (2003)</th><th>Borman & Brush (1993)</th><th>Shore et al. (1990)</th></tr>
<tr><td>Achievement motivation</td><td>Drive</td><td>Useful personal behavior</td><td rowspan="2">Performance style</td></tr>
<tr><td>Analysis</td><td>Problem solving</td><td>Technical activities and the mechanics of management</td></tr>
<tr><td>Interpersonal skills</td><td rowspan="2">Consideration and awareness of others</td><td rowspan="3">Interpersonal dealings and communication</td><td rowspan="4">Interpersonal style</td></tr>
<tr><td>Dealing with conflicts</td></tr>
<tr><td>Oral communication</td><td>Communication</td></tr>
<tr><td>Influencing others</td><td>Influencing others</td><td>Leadership</td></tr>
</table>

Table A7

Classification of the AC Dimensions Used in Sample 3 Into Broad Dimensions Based on Popular Taxonomies

<table>
<tr><th>Dimension</th><th>Arthur at al. (2003)</th><th>Borman & Brush (1993)</th><th>Shore et al. (1990)</th></tr>
<tr><td>Analytical skills</td><td>Problem solving</td><td rowspan="2">Technical activities and the mechanics of management</td><td rowspan="2">Performance style</td></tr>
<tr><td>Organizing and planning</td><td>Organizing and planning</td></tr>
<tr><td>Persuasiveness</td><td rowspan="2">Influencing others</td><td rowspan="2">Leadership</td><td rowspan="4">Interpersonal style</td></tr>
<tr><td>Assertiveness</td></tr>
<tr><td>Cooperation</td><td>Consideration and awareness of others</td><td rowspan="2">Interpersonal dealings and communication</td></tr>
<tr><td>Presentation skills</td><td>Communication</td></tr>
</table>

General Discussion

The present thesis aimed to contribute to the understanding of AC construct-related validity and to provide practical guidance in this regard. Thereby, it referred to different explanations for the findings on internal construct-related validity of ACs and also to an alternative, external construct-related validation approach for determining to what degree ACs measure the intended dimensions.

In this chapter, I will first summarize the main findings and contributions of the studies that were conducted for this thesis. Then, I will address strengths and limitations of this thesis. Finally, practical implications and directions for future research that can be deduced from each of the studies will be presented.

Main Findings and Contributions

The study presented in *Chapter 1* examined the trade-offs between two factors associated with assessor expertise, namely assessor training and assessor background on the one hand, and assessor team size on the other hand, in affecting rating accuracy in an AC exercise. These factors are all related to the costs of ACs and, therefore, it was of particular interest whether increasing assessor team size could compensate for missing assessor expertise and vice versa. Results revealed that increasing the size of the assessor team could only partially compensate for missing assessor training, in particular when assessors did not have a psychological background. However, increasing the size of the assessor team could compensate for using assessors with a suboptimal background – irrespective of whether assessors were trained or not. These findings imply that appropriate assessor training is essential for rating accuracy in ACs (see also Lievens, 2001a; Schleicher, Day, Mayes, & Riggio, 2002; Woehr & Huffcutt, 1994) because it cannot always be substituted for by aggregating ratings from multiple assessors. If no training is provided to assessors, increasing

the size of the assessor team can improve rating accuracy in an AC to some degree. However, in this case, it is recommended to use assessors with a psychological background. This is because assessors with a psychological background have less difficulty in differentiating between dimensions than managers who have a non-psychological background, for example (Lievens, 2001a; 2001b), and, therefore, increasing the size of the assessor team seems to be more effective in improving rating accuracy for the former than for the latter. Based on these findings, this study provided important practical guidance on how to weigh assessor expertise against the size of the assessor team so that rating accuracy can be ensured while keeping AC costs under control. As rating accuracy is connected to AC construct-related validity (Gaugler & Thornton, 1989; Lievens, 2001a; Melchers, Kleinmann, & Prinz, 2010; Schleicher et al., 2002), the results from this study are also relevant for AC construct-related validity.

Chapter 2 presented a study on the effects of exercise similarity on AC construct-related and criterion-related validity. By investigating AC construct-related and criterion-related validity simultaneously, we followed recent claims for a broad validation strategy (e.g., Lievens, 2009; Lievens, Dilchert, & Ones, 2009; Melchers & König, 2008; Woehr & Arthur, 2003). In doing so, we were able to answer the question of whether improvements in one aspect of validity are always paralleled by improvements in the other aspect as might be assumed according to the unitarian framework of validity (e.g., Binning & Barrett, 1989; Landy, 1986; Messick, 1995). In line with our expectations and with previous findings (e.g., Highhouse & Harris, 1993; Schneider & Schmitt, 1992), convergence between dimension ratings was higher when exercises were similar compared to when they were dissimilar, indicating that exercise similarity is beneficial for construct-related validity. However, criterion-related validity of ratings from similar and dissimilar exercises did not differ, indicating that exercise similarity does not impair criterion-related validity as might be expected (cf. Gaugler, Rosenthal, Thornton, & Bentson, 1987; Lievens et al., 2009). Taken

together, our findings suggest that interventions to improve one aspect of validity are not necessarily paralleled by improvements in the other aspect of validity. Thus, this study contributed to a greater understanding of the connection between the construct-related and criterion-related validity of ACs, which is also of importance from a practical perspective.

In *Chapter 3*, the focus was on the suggestion to relate AC overall dimension ratings to external evaluations of the same dimensions to find evidence for construct-related validity of ACs (cf. Arthur, Day, & Woehr, 2008; Neidig & Neidig, 1984; Reilly, Henry, & Smither, 1990; Rupp, Thornton, & Gibbons, 2008). Specifically, we examined whether this approach to external construct-related validation is successful in evidencing substantial dimension variance in dimension ratings from an AC. Consistent findings across three independent samples revealed that variance in AC overall dimension ratings and external dimension ratings can be ascribed primarily to source factors. Dimension factors were not found in the latent factor structure of AC overall dimension ratings and external dimension ratings, indicating that these ratings cannot be attributed to latent factors reflecting the targeted dimensions. Thus, this study showed that relating AC overall dimension ratings to external ratings of the same dimension does not provide evidence for AC construct-related validity as expected by some researchers (cf. Arthur et al., 2008; Neidig & Neidig, 1984; Reilly et al., 1990; Rupp et al., 2008). Rather, the results suggest that different sources have different perspectives on candidates' performance, which is in line with theoretical assumptions and findings from the multisource feedback domain (cf. Borman, 1974; 1977; Conway, 1996; Hoffman, Lance, Bynum, & Gentry, 2010; Woehr, Sheehan, & Bennett, 2005). In general, this study contributed towards a more comprehensive insight into the construct-related validity of ACs, and it offered some practical guidance concerning the use of dimension ratings and the use of information from different sources as a basis for personnel decisions and feedback to candidates.

Strengths and Limitations

Specific strengths and limitations of the studies conducted were discussed in *Chapters 1* to *3*. In the following, the major strengths and limitations of the presented research will be considered at a more general level.

One of the strengths of this thesis is that it offered a broad view on AC construct-related validity. In particular, *Chapters 1* to *3* addressed different explanations for the findings concerning the construct-related validity of ACs and moderators of internal construct-related validity that can be inferred from these explanations, respectively. The assumption that a lack of internal construct-related validity is due to biases on the side of assessors (cf. Lievens & Klimoski, 2001; Zedeck, 1986) formed the background of *Chapter 1*. Specifically, *Chapter 1* focused on assessor expertise as a moderator of AC construct-related validity and its effects on rating accuracy. In addition, increasing the size of the assessor team was considered to be a potential means to compensate for missing assessor expertise in improving rating accuracy. *Chapter 2* dealt with exercise similarity as a moderator of construct-related validity of ACs that can be inferred from the situational specificity hypothesis that has been offered as an alternative explanation for the findings on internal construct-related validity of ACs (e.g., Lance, Foster, Gentry, & Thoresen, 2004; Lance, Foster, Nemeth, Gentry, & Drollinger, 2007; Lance et al., 2000). Finally, based on arguments concerning the nature of ACs and the aforementioned situational specificity hypothesis*, Chapter 3* followed the suggestion to relate AC overall dimension ratings to external ratings of the same dimensions to demonstrate AC construct-related validity (cf. Arthur et al., 2008; Neidig & Neidig, 1984; Reilly et al., 1990; Rupp et al., 2008). Thus, *Chapters 1* to *3* offered different perspectives on the domain of AC construct-related validity.

The second strength of this thesis is that the studies presented in *Chapters 2* and *3* both reported AC construct-related and criterion-related validity. This is in contrast to most of the

previous studies in the AC domain that usually focused on only one aspect of validity and thus allowed only limited conclusions. By investigating construct-related and criterion-related validity simultaneously, we followed recent calls to use a broad validation strategy (e.g., Lievens, 2009; Lievens et al., 2009; Melchers & König, 2008; Woehr & Arthur, 2003). On the one hand, this allowed us to demonstrate that the ACs considered had criterion-related validity, meaning that they served their purpose of predicting job performance regardless of whether or not they had construct-related validity. On the other hand, we were able to examine whether a specific moderator of construct-related validity has parallel effects on criterion-related validity (*Chapter 2*). Thus, the broad validation strategy applied in this thesis contributed to the understanding of AC construct-related validity and its connection to criterion-related validity.

A further strength is that all studies conducted within this thesis considered recommendations concerning the design of ACs that can be found in the literature (e.g., International Task Force on Assessment Center Guidelines, 2009; Klimoski & Brickner, 1987; Lievens, 1998) as much as possible. For example, assessors were trained, and the number of dimensions that had to be observed simultaneously was limited. Thus, the ACs considered were comparable to operational ACs following common recommendations and, therefore, the results from our studies cannot be attributed to special characteristics of the ACs considered.

The fourth strength of this thesis is that it offered useful guidance concerning the design of ACs and thus provided a contribution for AC practice. The reported findings allowed conclusions concerning assessor training, assessor background, and the size of the assessor teams (*Chapter 1*). Furthermore, implications for the similarity of AC exercises (*Chapter 2*) and the use of dimension ratings (*Chapter 3*) were presented.

Besides the aforementioned strengths, this thesis also has a limitation that needs to be considered. Data for two of the presented studies (see *Chapters 1* and *2*) were exclusively gathered in a laboratory setting. This calls the external validity of the results obtained into question. However, in both studies, the exercises and dimensions used were comparable to those usually found in practice, and assessors were prepared for their rating task. Therefore, despite the laboratory setting, the assessors' task was representative of the rating task in the field and thus results should be generalizable. Moreover, in the study in which we used a simulated AC (see *Chapter 2*), candidates reported that they acted like they would in a real selection situation.

Practical Implications

The focus of this paragraph is on the major practical implications that can be derived from the studies presented in *Chapters 1* to *3*. The findings presented in *Chapter 1* suggest that some interventions to improve AC construct-related validity and rating accuracy, respectively, can (at least partially) compensate for each other. These findings imply that a few selected interventions might suffice to obtain accurate dimension ratings and that additional interventions might not bring further meaningful improvements in rating accuracy and AC construct-related validity, respectively. Therefore, AC users should carefully decide which interventions to improve rating accuracy and construct-related validity are most appropriate for a particular AC. Thereby they should consider the feasibility as well as the cost-effectiveness of these interventions to arrive at a reasonable decision.

According to the findings reported in *Chapter 2*, interventions to improve one aspect of validity do not necessarily improve other aspects of validity. Therefore, AC users would do better to design ACs so that they serve their specific purpose best. For example, when an AC serves to decide on candidates' developmental needs, then construct-related validity is of

particular importance. In this case, factors that have been found to moderate construct-related validity should be considered when designing the AC. However, taking moderators of criterion-related validity into account cannot be expected to improve construct-related validity, in spite of predictions from the unitarian framework of validity for a positive effect.

The results of the study presented in *Chapter 3* revealed that ACs provide a perspective on candidates' performance that differs from perspectives provided by other sources, such as customers or supervisors, for example. Thus, even though ratings from these other sources and from the AC do not seem to reflect a common set of dimensions, our results suggest that using ratings from an AC in addition to performance evaluations from other sources might provide a better basis for important personnel decisions than information from a single source. Therefore, ACs should be regarded as a supplement to other methods for performance evaluation.

Directions for Future Research

Finally, this thesis provides directions for future research. AC research has offered several explanations for the findings on internal construct-related validity of ACs and has identified a range of moderators of AC validity. Based on this, interventions to improve AC construct-related validity have been proposed. However, interaction effects of such interventions are relatively unknown to date. Future research should examine the interaction effects of different interventions to improve AC construct-related validity. This would allow answering questions such as what combination of interventions to reduce cognitive demands placed on assessors is useful or whether or not additional interventions have added value, for example. Thereby, of particular interest might be whether expensive interventions (e.g., using expert assessors) can be substituted by more cost-effective or feasible interventions (e.g., providing assessors with behavioral checklists).

Furthermore, most of the previous studies in the AC domain focused on only one aspect of validity. However, improvements in AC construct-related validity do not necessarily lead to parallel improvements in criterion-related validity as might be assumed based on the unitarian framework of validity (e.g., Binning & Barrett, 1989). Therefore, more research that uses a broad validation strategy is needed (cf. Lievens, 2009; Lievens et al., 2009; Melchers & König, 2008; Woehr & Arthur, 2003). That means future research should simultaneously focus on different aspects of validity, namely on construct-related, criterion-related, and also content-related validity of ACs, particularly when examining the effects of moderators of AC validity. This would contribute to the understanding of the connections between different aspects of AC validity, which is also of practical relevance. Specifically, AC users could be provided with information concerning simultaneous consequences of specific interventions to improve construct-related validity for content-related and criterion-related validity. For example, it might be interesting to determine whether reducing the number of dimensions that have to be observed at the same time has comparable effects on content-related and criterion-related validity as on construct-related validity of an AC (cf. Gaugler & Thornton, 1989).

On the one hand, this thesis focused on moderators of AC validity. On the other hand, the external construct-related validation approach for ACs was of particular interest. Thereby, the focus was on the relation between AC overall dimension ratings and ratings of the same dimensions provided by other sources that referred to performance on the job. An interesting direction for future research would be to consider other approaches to examine the external construct-related validity that have been given little attention so far. AC dimension ratings could be related to evaluations of the same dimensions that were gathered in other situations in which candidates were motivated to present themselves in their best light, such as an interview or another AC, for example. This or other similar approaches might be successful in evidencing that dimension ratings from ACs reflect performance on the purported dimensions.

References

Arthur, W., Jr., Day, E. A., & Woehr, D. J. (2008). Mend it, don't end it: An alternate view of assessment center construct-related validity evidence. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 105-111. doi:10.1111/j.1754-9434.2007.00019.x

Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology, 74*, 478-494. doi:10.1037/0021-9010.74.3.478

Borman, W. C. (1974). The rating of individuals in organizations: An alternate approach. *Organizational Behavior & Human Performance, 12*, 105-124. doi:10.1016/0030-5073(74)90040-3

Conway, J. M. (1996). Analysis and design of multitrait-multirater performance appraisal studies. *Journal of Management, 22*, 139-162. doi:10.1177/014920639602200106

Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., III, & Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology, 72*, 493-511. doi:10.1037/0021-9010.72.3.493

Gaugler, B. B., & Thornton, G. C., III. (1989). Number of assessment center dimensions as a determinant of assessor accuracy. *Journal of Applied Psychology, 74*, 611-618. doi:10.1037/0021-9010.74.4.611

Highhouse, S., & Harris, M. M. (1993). The measurement of assessment center situations: Bem's template matching technique for examining exercise similarity. *Journal of Applied Social Psychology, 23*, 140-155. doi:10.1111/j.1559-1816.1993.tb01057.x

Hoffman, B., Lance, C. E., Bynum, B., & Gentry, W. A. (2010). Rater source effects are alive and well after all. *Personnel Psychology, 63*, 119-151. doi:10.1111/j.1744-6570.2009.01164.x

International Task Force on Assessment Center Guidelines. (2009). Guidelines and ethical considerations for assessment center operations. *International Journal of Selection and Assessment, 17*, 243-253. doi:10.1111/j.1468-2389.2009.00467.x

Klimoski, R., & Brickner, M. (1987). Why do assessment centers work? The puzzle of assessment center validity. *Personnel Psychology, 40*, 243-260. doi:10.1111/j.1744-6570.1987.tb00603.x

Lance, C. E., Foster, M. R., Gentry, W. A., & Thoresen, J. D. (2004). Assessor cognitive processes in an operational assessment center. *Journal of Applied Psychology, 89*, 22-35. doi:10.1037/0021-9010.89.1.22

Lance, C. E., Foster, M. R., Nemeth, Y. M., Gentry, W. A., & Drollinger, S. (2007). Extending the nomological network of assessment center construct validity: Prediction of cross-situationally consistent and specific aspects of assessment center performance. *Human Performance, 20*, 345-362.

Lance, C. E., Newbolt, W. H., Gatewood, R. D., Foster, M. R., French, N. R., & Smith, D. E. (2000). Assessment center exercise factors represent cross-situational specificity, not method bias. *Human Performance, 13*, 323-353. doi:10.1207/S15327043HUP1304_1

Landy, F. J. (1986). Stamp collecting versus science: Validation as hypothesis testing. *American Psychologist, 41*, 1181-1192. doi:10.1037/0003-066X.41.11.1183

Lievens, F. (1998). Factors which improve the construct validity of assessment centers: A review. *International Journal of Selection and Assessment, 6*, 141-152. doi:10.1111/1468-2389.00085

Lievens, F. (2001a). Assessor training strategies and their effects on accuracy, interrater reliability, and discriminant validity. *Journal of Applied Psychology, 86*, 255-264. doi:10.1037/0021-9010.86.2.255

Lievens, F. (2001b). Assessors and use of assessment centre dimensions: A fresh look at a troubling issue. *Journal of Organizational Behavior, 22*, 203-221. doi:10.1002/job.65

Lievens, F. (2009). Assessment centres: A tale about dimensions, exercises, and dancing bears. *European Journal of Work and Organizational Psychology, 18*, 102-121. doi:10.1080/13594320802058997

Lievens, F., Dilchert, S., & Ones, D. S. (2009). The importance of exercise and dimension factors in assessment centers: Simultaneous examinations of construct-related and criterion-related validity. *Human Performance, 22*, 375-390. doi:10.1080/08959280903248310

Lievens, F., & Klimoski, R. J. (2001). Understanding the assessment center process: Where are we now? In C. L. Cooper & I. T. Robertson (Eds.), *International Review of Industrial and Organizational Psychology* (Vol. 16, pp. 245-286). Chichester, UK: Wiley.

Melchers, K. G., Kleinmann, M., & Prinz, M. A. (2010). Do assessors have too much on their plates? Rating quality and the number of simultaneously observed candidates in assessment center group discussions. *International Journal of Selection and Assessment, 18*, 329-341. doi:10.1111/j.1468-2389.2010.00516.x

Melchers, K. G., & König, C. J. (2008). It is not yet time to dismiss dimensions in assessment centers. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 125-127. doi:10.1111/j.1754-9434.2007.00023.x

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749. doi:10.1037/0003-066X.50.9.741

Neidig, R. D., & Neidig, P. J. (1984). Multiple assessment center exercises and job relatedness. *Journal of Applied Psychology, 69*, 182-186. doi:10.1037/0021-9010.69.1.182

Reilly, R. R., Henry, S., & Smither, J. W. (1990). An examination of the effects of using behavior checklists on the construct validity of assessment center dimensions. *Personnel Psychology, 43*, 71-84. doi:10.1111/j.1744-6570.1990.tb02006.x

Rupp, D. E., Thornton, G. C., III, & Gibbons, A. M. (2008). The construct validity of the assessment center method and usefulness of dimensions as focal constructs. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 116-120. doi:10.1111/j.1754-9434.2007.00021.x

Schleicher, D. J., Day, D. V., Mayes, B. T., & Riggio, R. E. (2002). A new frame for frame-of-reference training: Enhancing the construct validity of assessment centers. *Journal of Applied Psychology, 87*, 735-746. doi:10.1037/0021-9010.87.4.735

Schneider, J. R., & Schmitt, N. (1992). An exercise design approach to understanding assessment center dimension and exercise constructs. *Journal of Applied Psychology, 77*, 32-41. doi:10.1037/0021-9010.77.1.32

Woehr, D. J., & Arthur, W., Jr. (2003). The construct-related validity of assessment center ratings: A review and meta-analysis of the role of methodological factors. *Journal of Management, 29*, 231-258. doi:10.1177/014920630302900206

Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology, 67*, 189-205. doi:10.1111/j.2044-8325.1994.tb00562.x

Woehr, D. J., Sheehan, M., & Bennett, W., Jr. (2005). Assessing measurement equivalence across rating sources: A multitrait-multirater approach. *Journal of Applied Psychology, 90*, 592-600. doi:10.1037/0021-9010.90.3.592

Zedeck, S. (1986). A process analysis of the assessment center method. *Research in Organizational Behavior, 8*, 259-296.

Curriculum Vitae

Andreja Wirz-Rodella

Date of birth 02.02.1982

Place of origin Zürich, Switzerland

Education

2002 - 2008	Master of Science in Psychology Universität Zürich, Switzerland
2001 - 2002	Studies in Business Administration Universität Zürich, Switzerland
1994 - 2001	Swiss Matura (i.e., qualification for university entrance) Kantonsschule Oerlikon, Zürich, Switzerland

Professional Experience

5/2008 - present	Assistant/doctoral student at the Department of Work and Organizational Psychology (Prof. Dr. Martin Kleinmann), Universität Zürich
04/2007 - 08/2007	Internship in the field of human resource development, Kienbaum (Schweiz) AG, Zürich
03/2006 – 01/2007	Junior consultant, Pro Informatik GmbH, Zürich
06/2006	Assessor in a research project at the Department of Work and Organizational Psychology, Universität Zürich, Zürich
07/2005 - 10/2005	Clinical internship, Psychiatrische Klinik, Münsterlingen
09/2004 - 02/2006	Back office assistant, Pro Informatik GmbH, Zürich

Publications and Conference Presentations

Melchers, K. G., Wirz, A., & Kleinmann, M. (in press). Dimensions AND exercises: Theoretical background of mixed-model assessment centers. In D. J. R. Jackson, C. E. Lance, & B. J. Hoffman (Eds.), *The psychology of assessment centers.* New York: Routledge.

Wirz, A., Melchers, K. G., Schultheiss, S. & Kleinmann, M. (2011, September). *The effects of exercise similarity on construct-related and criterion-related validity of an assessment center*. Paper presented at the 12th Congress of the Swiss Psychological Society, Fribourg, Switzerland.

Wirz, A., Schultheiss, S., Melchers, K. G. & Kleinmann, M. (2011, September). *Der Einfluss der Übungsähnlichkeit auf die Konstrukt- und Kriteriumsvalidität eines Assessment Centers*. Vortrag an der 7. Tagung der Fachgruppe Arbeits-, Organisations- und Wirtschaftspsychologie der Deutschen Gesellschaft für Psychologie, Rostock, Germany.

Wirz, A., Melchers, K. G., Lievens, F., De Corte, W. & Kleinmann, M. (2010, September). *Auch viele Beobachter ersetzen fehlende Expertise nicht, um akkurate Beurteilungen zu erhalten*. Poster präsentiert am 47. Kongress der Deutschen Gesellschaft für Psychologie, Bremen, Germany.

Wirz, A., Melchers, K. G. & Kleinmann, M. (2009, September). *Kognitive Anforderungen und Expertise von Assessoren in Assessment Centern.* Vortrag an der 6. Tagung der Fachgruppe Arbeits- und Organisationspsychologie der Deutschen Gesellschaft für Psychologie, Wien, Austria.

Wirz, A., Melchers, K. G. & Kleinmann, M. (2009, April). *Do cognitive demands and assessors' expertise affect assessment center construct-related validity?* Poster presented at the 24th annual conference of the Society for Industrial and Organizational Psychology (SIOP), New Orleans, LA.